

Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins

Matthew Slattery,^{1,7,8} Todd Riley,^{2,3,7} Peng Liu,^{3,4,7} Namiko Abe,¹ Pilar Gomez-Alcala,^{2,5} Iris Dror,⁶ Tianyin Zhou,⁶ Remo Rohs,^{6,*} Barry Honig,^{3,4,*} Harmen J. Bussemaker,^{2,3,*} and Richard S. Mann^{1,*}

¹Department of Biochemistry and Molecular Biophysics, Columbia University, 701 West 168th Street, HHSC 1104, New York, NY 10032, USA

²Department of Biological Sciences, Columbia University, 1212 Amsterdam Avenue, New York, NY 10027, USA

³Center for Computational Biology and Bioinformatics, Columbia University, 1130 St. Nicholas Avenue, New York, NY 10032, USA

⁴Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Columbia University, 1130 St. Nicholas Avenue, New York, NY 10032, USA

⁵Department of Electrical Engineering, Columbia University, 500 West 120th Street, New York, NY 10027, USA

⁶Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, USA

⁷These authors contributed equally to this work

⁸Present address: Institute for Genomics and Systems Biology, University of Chicago, 900 East 57th Street KCBD 10115, Chicago, IL 60637, USA

*Correspondence: rohs@usc.edu (R.R.), bh6@columbia.edu (B.H.), hjb2004@columbia.edu (H.J.B.), rsm10@columbia.edu (R.S.M.)

DOI 10.1016/j.cell.2011.10.053

SUMMARY

Members of transcription factor families typically have similar DNA binding specificities yet execute unique functions in vivo. Transcription factors often bind DNA as multiprotein complexes, raising the possibility that complex formation might modify their DNA binding specificities. To test this hypothesis, we developed an experimental and computational platform, SELEX-seq, that can be used to determine the relative affinities to any DNA sequence for any transcription factor complex. Applying this method to all eight *Drosophila* Hox proteins, we show that they obtain novel recognition properties when they bind DNA with the dimeric cofactor Extradenticle-Homothorax (Exd). Exd-Hox specificities group into three main classes that obey Hox gene collinearity rules and DNA structure predictions suggest that anterior and posterior Hox proteins prefer DNA sequences with distinct minor groove topographies. Together, these data suggest that emergent DNA recognition properties revealed by interactions with cofactors contribute to transcription factor specificities in vivo.

INTRODUCTION

Gene regulatory information is encoded in genomic DNA sequences and interpreted by transcription factors that bind to specific sequences. Although the in vitro binding properties of transcription factors have been studied for many years, it has proven notoriously difficult to predict in vivo genomic binding

from in vitro sequence specificity. Whether or not a predicted binding site is occupied in vivo depends strongly on sequence and chromatin context as well as cell type (Gaulton et al., 2010; Guertin and Lis, 2010; Kaplan et al., 2011). While the amount of genome-wide binding varies greatly between transcription factors, typically only a small fraction of a transcription factor's preferred DNA sequences are occupied in vivo.

What makes in vivo binding more specific than in vitro binding? One possible answer is that the organization of the chromatin—for example, the position of nucleosomes—limits access to transcription factor binding sites (Wunderlich and Mirny, 2009). A second explanation has its root in the combinatorial nature of gene regulation. Unlike individual transcription factors, complexes of interacting factors bind cooperatively to genomic regions that contain a favorable configuration of binding sites (Johnson, 1995). These mechanisms, however, are unlikely to be sufficient to account for the transcription factor specificities observed in vivo. In particular, confounding the issue of specificity is that most transcription factors are members of protein families that have very similar DNA binding domains with similar recognition properties. For example, in the mouse there are nineteen T-box factors that can bind to variations of the sequence TCACACC, 39 Hox family homeodomain proteins that bind to AT-rich binding sites, and nearly 60 basic helix-loop-helix (bHLH) factors, most of which bind to the DNA sequence CACGTG known as the “E-box” (Berger et al., 2008; Conlon et al., 2001; Jones, 2004; Noyes et al., 2008). Despite overlapping binding specificities, these factors carry out distinct functions in vivo (Alexander et al., 2009; Cao et al., 2010; Naiche et al., 2005; Pearson et al., 2005). Although some specificity is derived from the cell type specific expression of individual family members, the fundamental question of how they recognize distinct binding sites and regulate unique sets of target genes in vivo remains unsolved.

Although members of the same transcription factor family typically have very similar DNA binding domains these domains are rarely identical, raising the possibility that small differences in protein sequence could lead to significant differences in binding specificity. However, when assayed *in vitro*, using either classical or high-throughput methods, different members of the same protein family generally do not show large differences in binding specificity. For example, in *Drosophila* more than 50 homeodomain proteins bind to the six-base-pair sequences TAATTG and TAATTA, despite differences in their DNA binding domains (Berger et al., 2008; Noyes et al., 2008). On the other hand, subtle differences in homeodomain sequences, and transcription factor sequences in general, are often conserved across vast evolutionary distances, arguing that these differences are functionally important. The eight Hox paralogs in *Drosophila*, for instance, which execute distinct functions *in vivo*, each have recognizable orthologs in both vertebrates and other invertebrates. Hox orthologs can be recognized not only by their protein sequences but also from the order in which they are expressed along an animal's anteroposterior (AP) axis (Hueber et al., 2010). Moreover, orthologous Hox proteins often have conserved functions when expressed in a heterologous species (Lutz et al., 1996; McGinnis et al., 1990; Zhao et al., 1993). These observations suggest that sequence differences between related transcription factors, although evolutionarily conserved and functionally relevant, are not typically reflected in differences in their DNA binding preferences.

There are two plausible solutions to this paradox. One is that some of the sequence differences between related transcription factors do not play a role in DNA binding, but instead affect their ability to repress or activate their target genes. Several examples of this so-called "activity regulation" have been described, and suggest that the ability to recruit different coactivators or corepressors may be used to diversify transcription factor function (Gebelein et al., 2004; Joshi et al., 2010; Li and McGinnis, 1999; Taghli-Lamalle et al., 2007). An alternative mechanism, which we refer to here as "latent specificity," is that differences in the amino acid sequences of transcription factors within the same structural family may only impact DNA recognition when these factors bind with cofactors. This mechanism is distinct from conventional cooperativity, in which binding energetics are affected by the presence of a cofactor but nucleotide sequence specificity is not. By contrast, in latent specificity there is a cofactor-induced change in DNA recognition. For example, as shown by X-ray crystallography, the *Drosophila* Hox protein Sex combs reduced (Scr) has distinct DNA recognition properties when it binds as a heterodimer with its cofactor Extradenticle (Exd) (Joshi et al., 2007). By directly binding a Hox peptide known as the "YPWM" motif, Exd helps to position the N-terminal arm of Scr's homeodomain so that it can recognize a sequence-dependent narrow minor groove in its DNA binding site. The binding to narrow minor grooves, typically by Arg residues, is an example of the widely used mechanism of DNA shape recognition (Rohs et al., 2009). Although Exd and its mammalian orthologs Pbx1-3 can heterodimerize with all Hox family members, and differences in DNA sequence preferences for Exd-Hox complexes have been reported (Chan et al., 1994; Chang et al., 1996; Lu and Kamps, 1997; Mann and Chan,

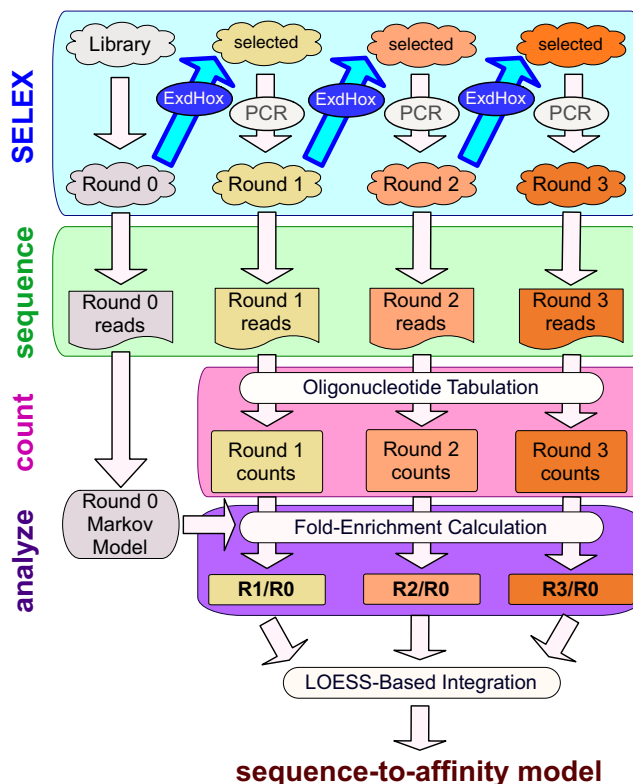


Figure 1. Overview of SELEX-seq

The starting point is a pool of synthesized DNA oligonucleotides containing a region of 16 random base pairs. This random pool is made double stranded and then sequenced using Illumina sequencing, resulting in a set of R0 reads. EMSAs are performed on the random pool and DNA molecules bound to Exd-Hox heterodimers are isolated and amplified by PCR. This enriched pool (R1) is sequenced. The affinity-based selection step is repeated multiple times. To accurately parameterize the sequence biases in R0, a Markov model is constructed. Relative fold-enrichments associated with the affinity-based selection step are calculated for all 12-mers. Information from earlier and later rounds of selection is combined using LOESS regression to estimate the relative binding affinity for each 12-mer with an optimal trade-off between accuracy and precision. See also Figure S1 and Tables S1 and S2.

1996), the degree to which the assembly of multi-protein complexes influences binding specificity has not been systematically analyzed for Hox proteins, or for any transcription factor family.

Here, we describe a high-throughput and systematic approach that demonstrates that complex formation between Hox factors and Exd uncovers latent DNA binding specificities that are only revealed upon heterodimerization. To do this, we combined Systematic Evolution of Ligands by Exponential Enrichment (Tuerk and Gold, 1990) with massively parallel sequencing (SELEX-seq) (Figure 1). The depth of the sequence information, combined with a biophysical model of the SELEX-seq data, allows us to calculate the relative affinity for any DNA sequence. We apply this method to all eight *Drosophila* Hox proteins in complex with the same cofactor, Exd. By analyzing the enrichment of oligonucleotides through several rounds of selection, we find that all Exd-Hox heterodimers prefer to bind the sequence GAYNNAY (where Y = T or C) and that the familiar

preference of Hox proteins for TAAT sequences no longer dominates. Different Exd-Hox heterodimers exhibit strong preferences for distinct subsets of this generalized binding site, leading to a unique binding “fingerprint” for each Exd-Hox complex. These results suggest that members of transcription factor families achieve specificity in part by forming complexes that modify their DNA recognition properties in precise ways.

RESULTS

Overview of SELEX-Seq

Our strategy for characterizing the DNA binding specificities of each of the eight *Drosophila* Hox proteins in complex with Exd is summarized in Figure 1 (see also Figure S1 available online). All eight Hox proteins (Labial, Lab; Proboscipedia, Pb; Deformed, Dfd; Scr; Antennapedia, Antp; two different isoforms of Ultrabithorax, Ubx1a and Ubx1Va; Abdominal-A, Abd-A; and Abdominal-B, Abd-B), containing all known Exd interaction motifs, were expressed in and purified from *E. coli* (Figure S2A and Experimental Procedures). Full-length Exd was purified bound to the “Homothorax-Meis” (HM) domain of Homothorax (Hth). The HM domain is similar to a naturally occurring isoform of Hth that is necessary and sufficient for Exd nuclear localization and sufficient to carry out most Hox-dependent functions of *hth* during embryogenesis (Noro et al., 2006; Ryoo et al., 1999). Thus, to optimize cooperativity and mimic in vivo function, the protein complexes used in the SELEX experiments were HM-Exd-Hox trimers, which we refer to here as Exd-Hox complexes.

Our in vitro selection experiments started with a pool of double-stranded DNA oligomers consisting of 16 random base pairs flanked by sequences needed for PCR amplification and sequencing on the Illumina platform (see Table S1 and Experimental Procedures). HM-Exd-Hox-DNA complexes were isolated by electrophoretic mobility shift assays (EMSAs) to ensure that oligomers bound by Exd-Hox (as opposed to Hox or Exd monomers) were selected. For each Hox protein at least three rounds of selection were performed (R1 to R3). The pool of selected DNA molecules was sequenced after each round (Figure 1).

Computational Inference of Sequence-to-Affinity Tables

The procedure for quantifying the DNA binding specificities of each Exd-Hox complex relies on analyzing data from the first few rounds of selection, thus avoiding overselection and allowing the discovery of binding sites covering the full range of affinities (Figure 1; Experimental Procedures). It was important to characterize the initial library, R0, as it had biases in sequence composition (Figure S1A). In addition, because the number of DNA molecules sequenced in each round ($\sim 10^7$ Illumina reads) is much smaller than the number of distinct DNA molecules in R0 ($> 10^9$), many sequences that contain a specific Exd-Hox binding site were not sequenced in R0. However a fifth-order Markov model accurately predicts the relative frequency of all 16-mers in R0 (Figure S1A), allowing us to interpret the later rounds.

To determine the number of consecutive base pairs that must be specified to fully capture Exd-Hox binding specificities we

tabulated the fraction of reads containing each DNA subsequence of a given length. Treating this table as a probability distribution, we computed, for each oligomer length, the information gain (Kullback-Leibler divergence of R2 relative to R0) experienced by the pool during selection. 12-mers were optimal for capturing the sequence specificity of Exd-Hox complexes (Figure 2A). Relative affinities for each 12-mer were computed by calculating the round-to-round enrichments. The most highly represented 12-mers were sequenced $\sim 10^2$ - 10^3 times in R1 and $\sim 10^4$ - 10^5 times in R2 (Table S2). Thus, while the sampling error can be high in R1 ($\sim 10\%$), the precision with which relative affinities can be quantified is dramatically improved in R2. Integrating the affinity estimates from the R1 versus R0 and R2 versus R0 comparisons using a LOESS-based regression procedure (Figure 2B and Figures S1C and S1D) yielded a sequence-to-affinity table that combines the higher accuracy of R1 with the higher precision of R2 (see Table S3 for an abbreviated list and <http://bussemakerlab.org/papers/SELEXseq2011/> for the full list of 12-mers).

Identification of Exd-Hox Binding Site Variants

To discover all possible binding sites for each Exd-Hox complex in an unbiased manner, we created affinity tables for all 8-mers using the same methodology as for 12-mers. We found the most favored 8-mer to be TGATTGAT (preferred by Exd-Lab and Exd-Pb), TGATTAAT (preferred by Exd-Dfd and Exd-Scr), or TGATTTAT (preferred by Exd-Antp, Exd-Ubx1a, Exd-Ubx1Va, Exd-AbdA, and Exd-AbdB) (Figure 2C). We refer to these as green, blue, and red binding sites, respectively. These three groups of Hox proteins define three main specificity classes, 1 to 3, which we further refine below. As expected, for each Exd-Hox complex the fraction of DNA molecules in the pool containing the optimal “color” 8-mer increases monotonically with the number of rounds; other motifs become enriched at lower levels in the early rounds, but are outcompeted by the optimal motif in subsequent rounds (Figure S2B).

To identify the complete set of sequences that Exd-Hox complexes are capable of binding, we performed a systematic and unbiased iterative selection of additional 8-mers in decreasing order of relative enrichment. This yielded seven additional motifs with a relative affinity above 25% (compared to one of the top three). We refer to these as the magenta, black, cyan, light green, orange, yellow, and purple motifs (Figure 2C). Each of the ten motif variants fits the consensus TGAYNNAY. With the exception of Exd-Pb, which selected a high percentage of sequences containing two Pb monomer sites, the large majority of selected sequences fit this consensus (Table S4 and Figure S3A). Depending on the Hox protein, 1%–5% of the selected 12-mers contain Exd-Exd dimer binding sites, which were confirmed by carrying out SELEX-seq using only HM-Exd, in the absence of any Hox protein. Although Exd-Exd-DNA complexes have a mobility in EMSAs that is similar to Exd-Hox heterodimers (Figure S2C), sequences selected by Exd-Exd dimers fit consensus sequences that are distinct from Exd-Hox binding sites and were therefore computationally removed from the Exd-Hox data sets (Table S4 and Experimental Procedures). We also confirmed that the SELEX-seq platform

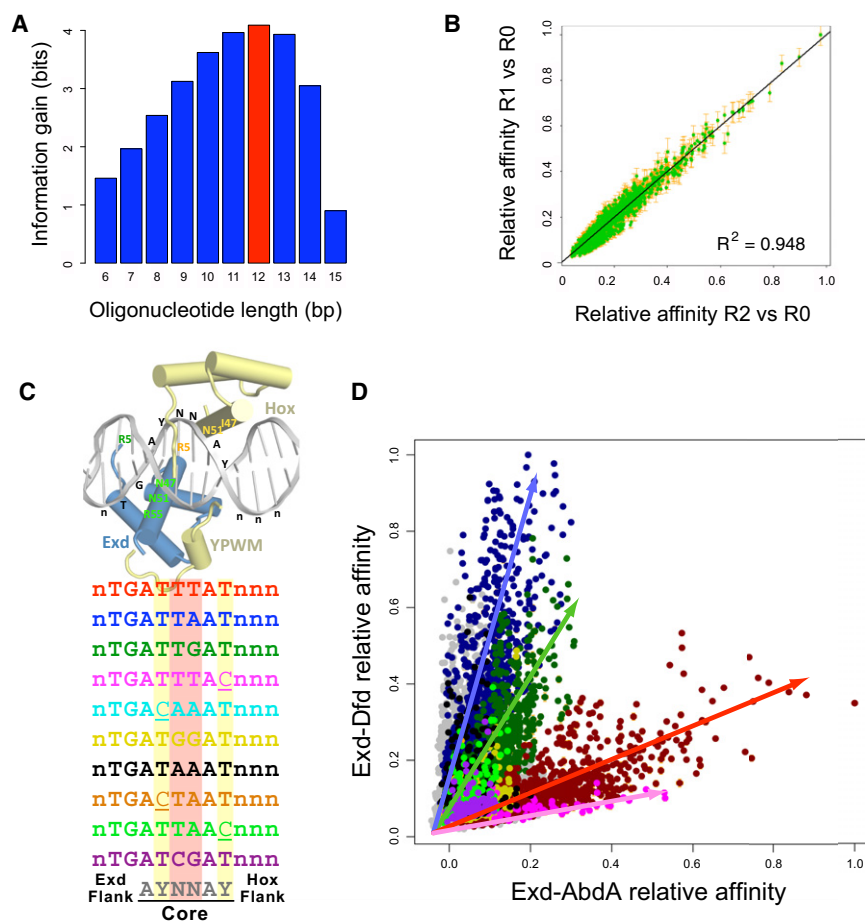


Figure 2. Multiple Core Sequences Support DNA Recognition by Exd-Hox Dimers

(A) Information gain (Kullback-Leibler divergence) associated with two rounds of affinity-based selection as a function of oligonucleotide length.

(B) Direct comparison between 12-mer affinities estimated as relative R0 \Rightarrow R1 enrichments and R0 \Rightarrow R2 enrichments corrected for nonlinear bias using LOESS regression. The error bars denote the standard error in the estimate of the relative affinity as calculated based on Poisson statistics (see [Extended Experimental Procedures](#)).

(C) Systematic discovery of Exd-Hox core binding motifs based on iterative selection of core motifs that are the most enriched after one round of selection. The most enriched sequences for any Exd-Hox contain one of the three primary motifs TGATTDAT (red, blue, green). Secondary motifs supporting a relative binding affinity of at least 25% all fit the consensus WRAYNNAY. The underlined base pairs indicate where Asn51 of the Exd and Hox homeodomains contacts the DNA, respectively. The IUPAC symbols “W” denotes A or T, “R” denotes A or G, “Y” denotes C or T, and “D” denotes not C.

(D) Scatter plot showing a direct comparison of the DNA binding preferences of Exd-Dfd and Exd-AbdA. Each point in the plot represents a unique 12-mer and is color-coded according to the core hexamer it contains; all possible 12-mers for which relative affinities could be determined are plotted. The error bars denote the standard error in the estimate of the relative affinity as calculated based on Poisson statistics (see [Extended Experimental Procedures](#)). The multiple diagonals with distinct slopes (arrows) indicate different relative preferences for the two dimers. The identities of the flanks modulate the binding affinity (distance from the origin). See also [Figure S2](#) and [Table S3](#).

generated reproducible relative affinities by comparing the data for two independent experiments for Exd-Scr ($R^2 = .929$; [Figure S2D](#)).

Modular Architecture of the Exd-Hox Binding Site

The highest-affinity 12-mer for each Exd-Hox dimer (cf. [Table S2](#)) has the form ATGATTDATNNN (where D = G, A, or T). As seen in X-ray structures, the first and second underlined A:T base pairs are contacted by Asn51 of the α 3 recognition helices of the Exd and Hox homeodomains in the major groove, respectively ([Joshi et al., 2007](#); [LaRonde-LeBlanc and Wolberger, 2003](#); [Passner et al., 1999](#); [Piper et al., 1999](#)) ([Figure 2C](#)). We refer to positions 1–3 and 10–12 as the Exd and Hox flanks, respectively, because the nucleotides in these positions only contact one of these two proteins. The central hexamer (positions 4–9), defined as the core motif, makes major and minor groove contacts with both Exd and Hox (see [Figures 2C](#) and discussion below).

The tripartite architecture (Exd flank – core hexamer – Hox flank) of the Exd-Hox binding site is helpful in interpreting our data. For example, plots of the relative binding affinities of Exd-Dfd versus Exd-AbdA for all 12-mers shows four distinct diagonals that correspond to different color core motifs. Dodecamers containing the magenta or red core hexamers are

preferred by Exd-AbdA, while those containing the blue or green core are preferred by Exd-Dfd ([Figure 2D](#)). This plot and similar plots (see below) suggest that the identity of the AYNNAY core of the Exd-Hox binding site is the primary determinant of binding preference for each Exd-Hox complex. The sequences flanking this core tune the affinity of the binding site. In comparative specificity plots such as this, differences in affinity due to different flanking sequences correspond to their distance from the origin, while differences in specificity correspond to distinct slopes.

Core Binding Site Preferences Differ between Exd-Hox Complexes

[Figure 3A](#) shows the distribution of 12-mer affinities partitioned by Hox identity and core motif color. Although we were unable to detect large differences in preference between the two Ubx isoforms (Ia and IVa), the other Exd-Hox complexes have a characteristic affinity “fingerprint” across the ten core motifs. For example, Exd-Lab and Exd-Pb are unique in that they do not bind well to the red sequences, while only Exd-Ubx fails to bind to yellow sequences. Similarly, only the abdominal Hox proteins (AbdB, AbdA, Ubx) bind with high relative affinity to the magenta sequences, while only Dfd and Scr bind well to the black and cyan sequences ([Figure 3A](#)).

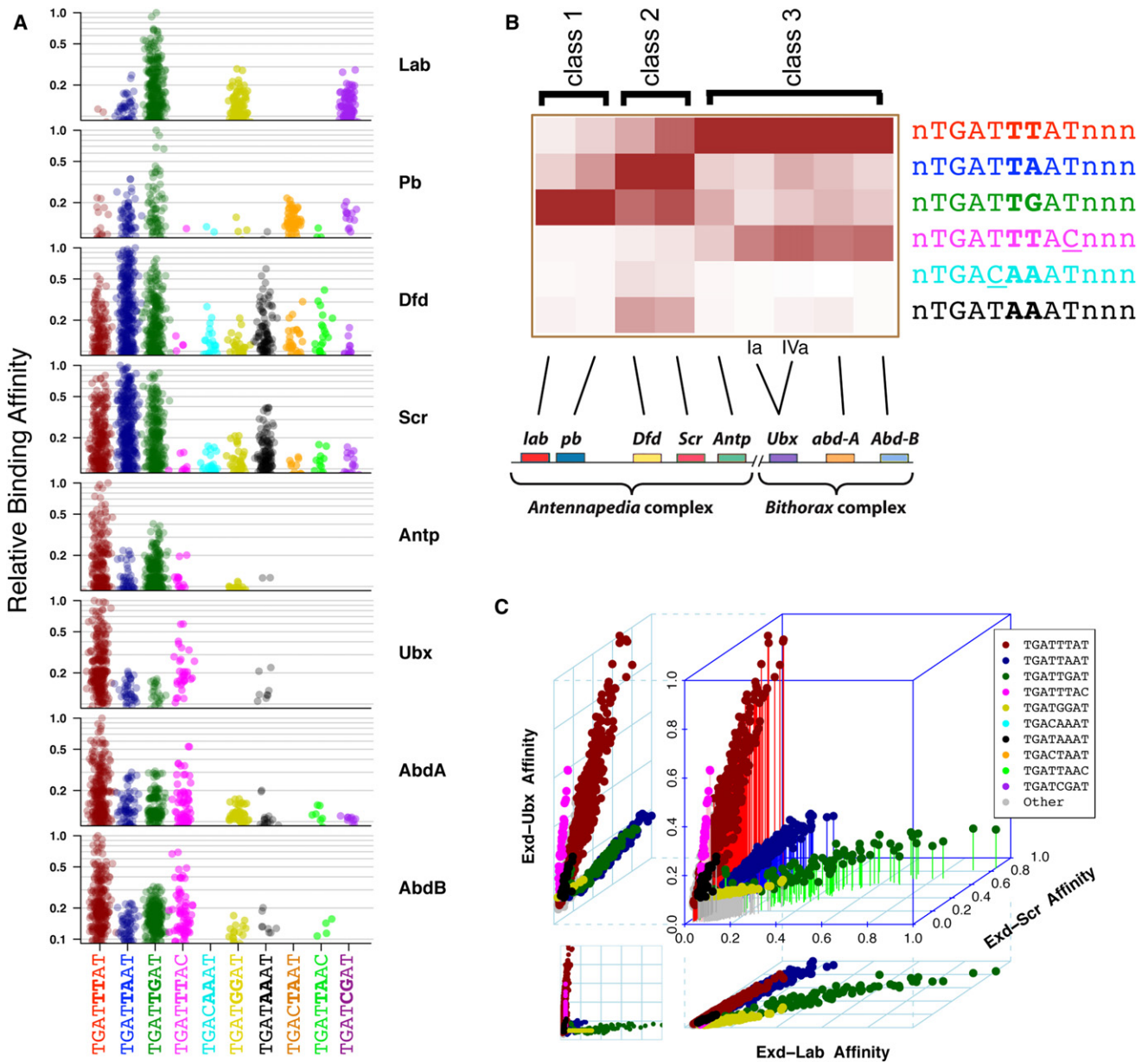


Figure 3. Exd-Hox Heterodimers Can Be Distinguished Based on Their DNA Specificity Fingerprints

(A) Strip charts (with arbitrary horizontal displacement) showing the distribution of relative affinities across all 12-mers for each Exd-Hox dimer. (B) Heat map of the Exd-Hox dimers based on the maximum relative affinity in each core motif class defines three major specificity classes, 1 to 3. The clustering is consistent with the linear ordering of the Hox genes along the chromosome. (C) Three-dimensional scatter plot comparing representative Exd-Hox complexes from each major specificity class. Two-dimensional projections for each pairwise comparison are shown. Color-coding is according to Figure 2C. See also Figure S3 and Table S4.

Representing these data in the form of a heat map reinforces the existence of three classes of binding site preferences that follow from the initially defined red, blue, and green motifs (Figure 3B). Remarkably, this classification is compatible with the order in which the Hox genes are positioned along the chromosome and with their expression domains along the anterior-posterior axis during *Drosophila* embryogenesis. Specificity class 2 proteins (Scr and Dfd) are the most promiscuous binders,

while the proteins in the other two classes are more selective. Antp differs from the other class 3 proteins, implying that this class should be divided into 3a (consisting of Antp), and 3b (consisting of Ubx, AbdA, and AbdB). The three specificity classes can also be visualized by three-dimensional comparisons of the binding site preferences for representative Exd-Hox dimers from each specificity class (Figure 3C), as well as by more traditional sequence logos (Figure S3A).

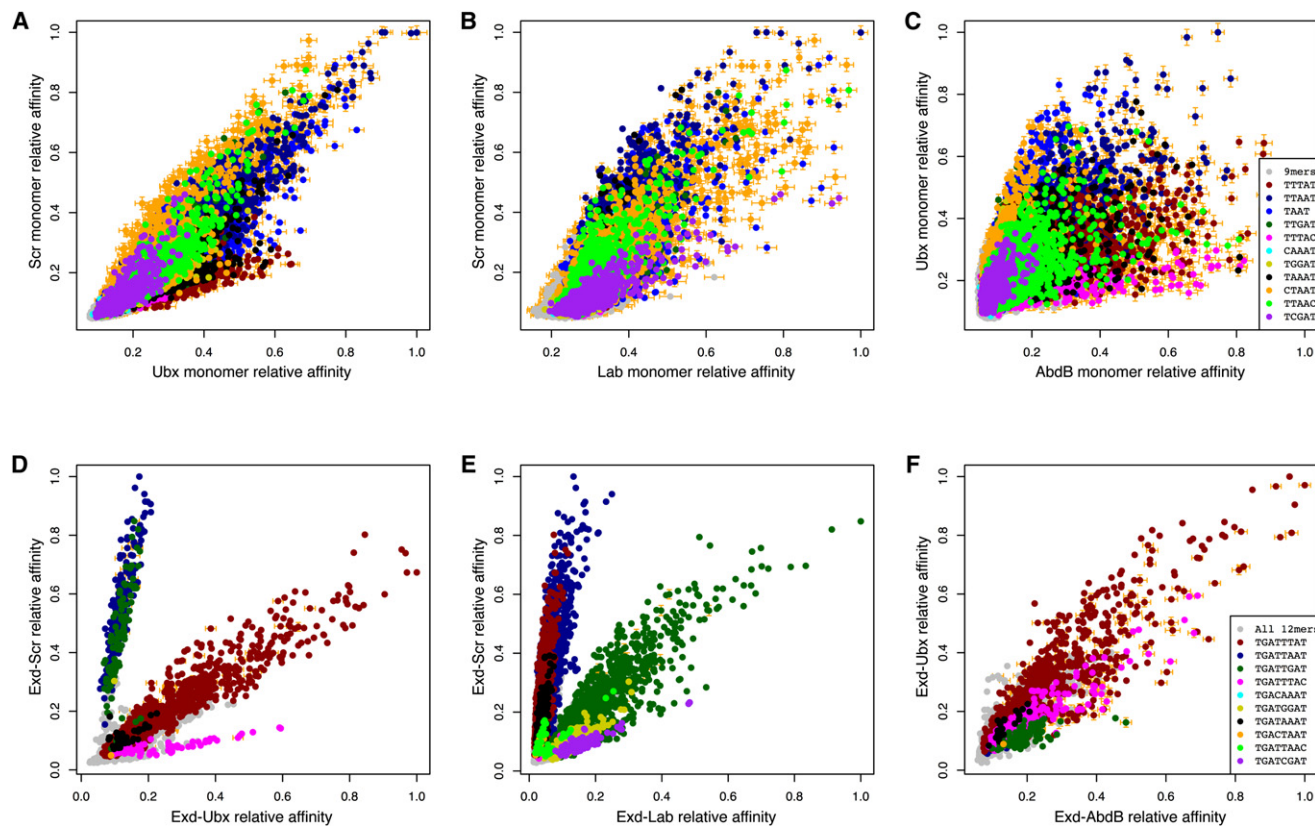


Figure 4. Heterodimerization with Exd Elicits Novel Binding Specificities

(A–C) Comparative specificity plots for monomeric Hox proteins showing relative affinities for all 9-mers. Comparing Scr versus Ubx (A) and Scr versus Lab (B) shows that there are only small differences in binding preference. Comparing Ubx versus AbdB (C) reveals that these two Hox proteins have both shared (e.g., light green) and distinct (e.g., orange for Ubx and magenta for AbdB) binding preferences. The error bars denote the standard error in the estimate of the relative affinity as calculated based on Poisson statistics (see [Extended Experimental Procedures](#)).

(D–F) Comparative specificity plots for Exd-Hox dimers showing relative affinities for all 12-mers. Comparing Exd-Scr versus Exd-Ubx (D) and Exd-Scr versus Exd-Lab (E) reveals differences in binding preference not observed for the corresponding monomer comparisons. Exd-Ubx versus Exd-AbdB (F) reveals a convergence of binding preference for red and magenta binding sites. The error bars denote the standard error in the estimate of the relative affinity as calculated based on Poisson statistics (see [Extended Experimental Procedures](#)). See also [Figure S3](#).

Unique Hox DNA Binding Preferences Are Revealed upon Heterodimerization with Exd

Previous work suggested that Exd allows Scr to bind DNA with greater specificity than it does as a monomer (Joshi et al., 2007). Using the SELEX-seq platform described here we tested this “latent specificity” hypothesis on a global scale, by comparing the specificities of four monomeric Hox proteins with the specificities of the same Hox proteins complexed with Exd. In all cases the Hox specificities are modified in the presence of Exd. Two pairwise comparisons of monomeric Hox binding preferences (Scr versus Labial and Scr versus Ubx) reveal the general tendency for all three of these Hox proteins to select sequences containing a TAAT, the motif that is traditionally associated with Hox binding sites (Figures 4A and 4B and [Figure S3B](#)). Although some modest preferences are observed (for example, Ubx prefers TTTAT more than Scr, [Figure 4A](#)), the monomeric specificities are not sufficient to distinguish between these Hox proteins, consistent with previous studies (Berger et al., 2008; Noyes et al., 2008). In contrast, when the

DNA binding preferences for the same Hox proteins are compared as complexes with Exd, a high degree of specificity is observed (Figures 4D and 4E). While red binding sites are bound well by both Exd-Scr and Exd-Ubx, the blue and green sites are bound more strongly by Exd-Scr than by Exd-Ubx. Conversely, the magenta site is bound more strongly by Exd-Ubx than by Exd-Scr ([Figure 4D](#)). Similarly, in the presence of Exd the specificities of Scr and Lab are readily distinguished, while the corresponding monomeric specificities are largely overlapping ([Figure 4B and 4E](#)).

Comparisons between AbdB and Ubx reveal a different type of Exd-dependent change in DNA binding specificity. AbdB’s binding site preferences as a monomer differ from those of the other seven Hox monomers (Berger et al., 2008; Noyes et al., 2008) ([Figure S3B](#)). Comparing the specificities of Ubx and AbdB monomers, for example, reveals that these two Hox proteins have both common and unique binding site preferences ([Figure 4C](#)). In contrast, the specificities of Exd-Ubx and Exd-AbdB are very similar; both prefer red and magenta binding sites

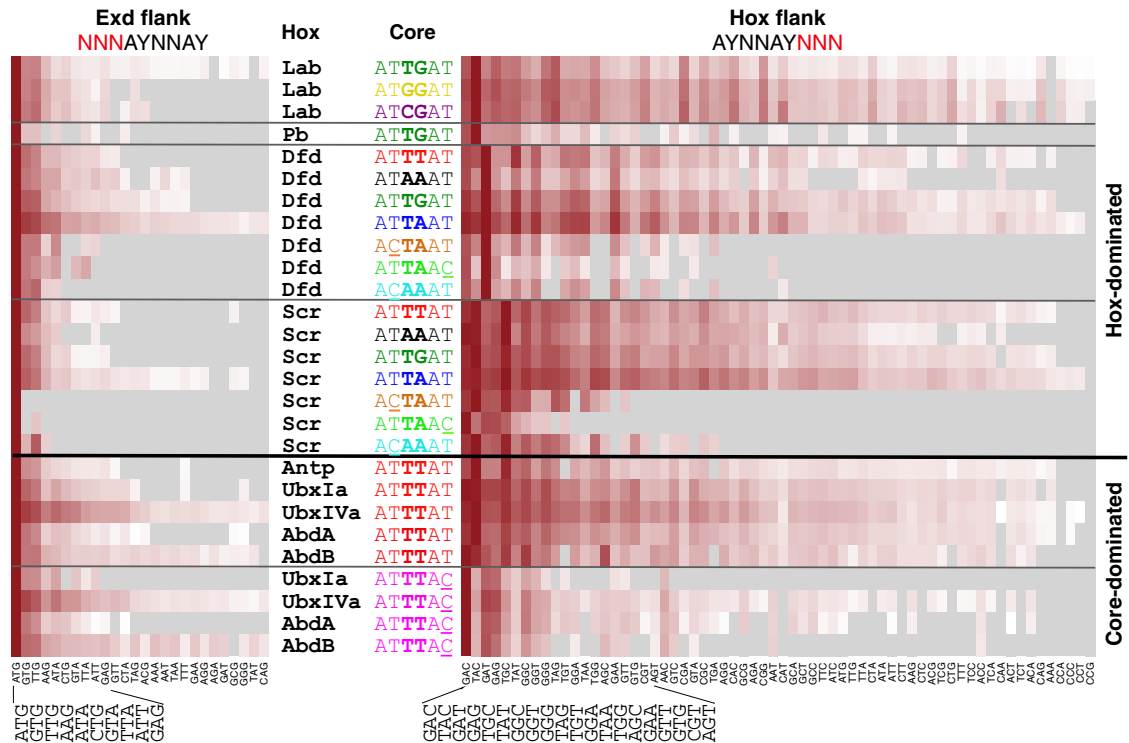


Figure 5. Modulation of Affinity and Specificity by the Exd and Hox Flanks

The relative affinities of all possible trinucleotides for both the Exd flank (left) and Hox flank (right) were analyzed in terms of their sequence context (Hox protein identity and core motif color). The number of trinucleotides displayed for the Exd flank was truncated because of the nearly complete absence of any binding for the less-preferred sequences. Preferences for the Hox flank depend on both the identity of the Hox protein (above the black line) and of the core motif (below the black line). Hox flank preference is dominated by Hox identity for class 1 and 2 Hox proteins, while it is dominated by core motif identity for class 3 Hox proteins. Gray positions denote sequences with < 100 counts (leading to relative errors greater than 10%), and have affinities less than the lightest colored cell for a given row.

(Figure 4F). Thus, in this case heterodimerization with Exd causes the specificities of these two class 3 Hox proteins to converge. It is also noteworthy that some monomer preferences (such as for the light green motif, TTAAC) are not observed for the corresponding heterodimers (compare Figure 4A–4C with Figure 4D–4F). Together, these findings demonstrate that Hox proteins have distinct DNA recognition properties as monomers compared to when they bind as heterodimers with Exd.

The Contribution of Exd and Hox Flanks to Binding Site Preferences

The above analysis demonstrates that differences in binding preference by different Exd-Hox dimers depend on which core motif (AYNNAY) is present. Previous work on monomeric homeodomain specificities emphasized a role for the $\alpha 3$ recognition helix in modulating DNA binding specificity (Berger et al., 2008; Gehring et al., 1994; Hanes and Brent, 1989; Mann, 1995; Noyes et al., 2008; Treisman et al., 1989). In Exd-Hox complexes, the Hox recognition helix contacts the Hox flank, as well as the core hexamer (Figure 2C). To investigate the contribution of these flanking base pairs to binding site preference in the context of Exd-Hox heterodimers, we analyzed the relative binding affinities for all trinucleotides in the Hox flank for each Exd-Hox dimer bound to its preferred core motifs. Similarly, we analyzed

the binding affinities for all possible trinucleotides in the Exd flank. Although a few alternative trinucleotides are tolerated, there is a Hox-independent preference for ATG, GTG, or TTG in the Exd flank (Figure 5). This preference makes sense as the Exd flank is contacted by Exd amino acid side chains (e.g., Arg5 in the minor groove and Arg55 in the major groove (Passner et al., 1999; Piper et al., 1999)), which are common to all Exd-Hox complexes (Figure 2C).

The situation is more complex for the Hox flank. In general, we find that the Hox flank preferences for Exd-Hox complexes are similar to the preferences for the equivalent base pairs in Hox monomer binding sites (Figure 5). However, which Hox flank is preferred depends in some cases on the Hox protein, while in other cases it depends on the core motif. For example, Dfd and Scr have different Hox flank preferences, regardless of which core motif is present, while the magenta and red sequences have Hox flank preferences that are independent of Hox protein identity (Figure 5). When Exd-AbdA binds to a red core motif it prefers the Hox flank TAC, but it prefers the Hox flanks GAC and GAT when binding to a magenta motif. Thus, in addition to revealing unique Hox flank preferences for the different Exd-Hox complexes, these data suggest there are mutual dependencies between the different parts of the Exd-Hox binding site (mostly between core motif and Hox flank).

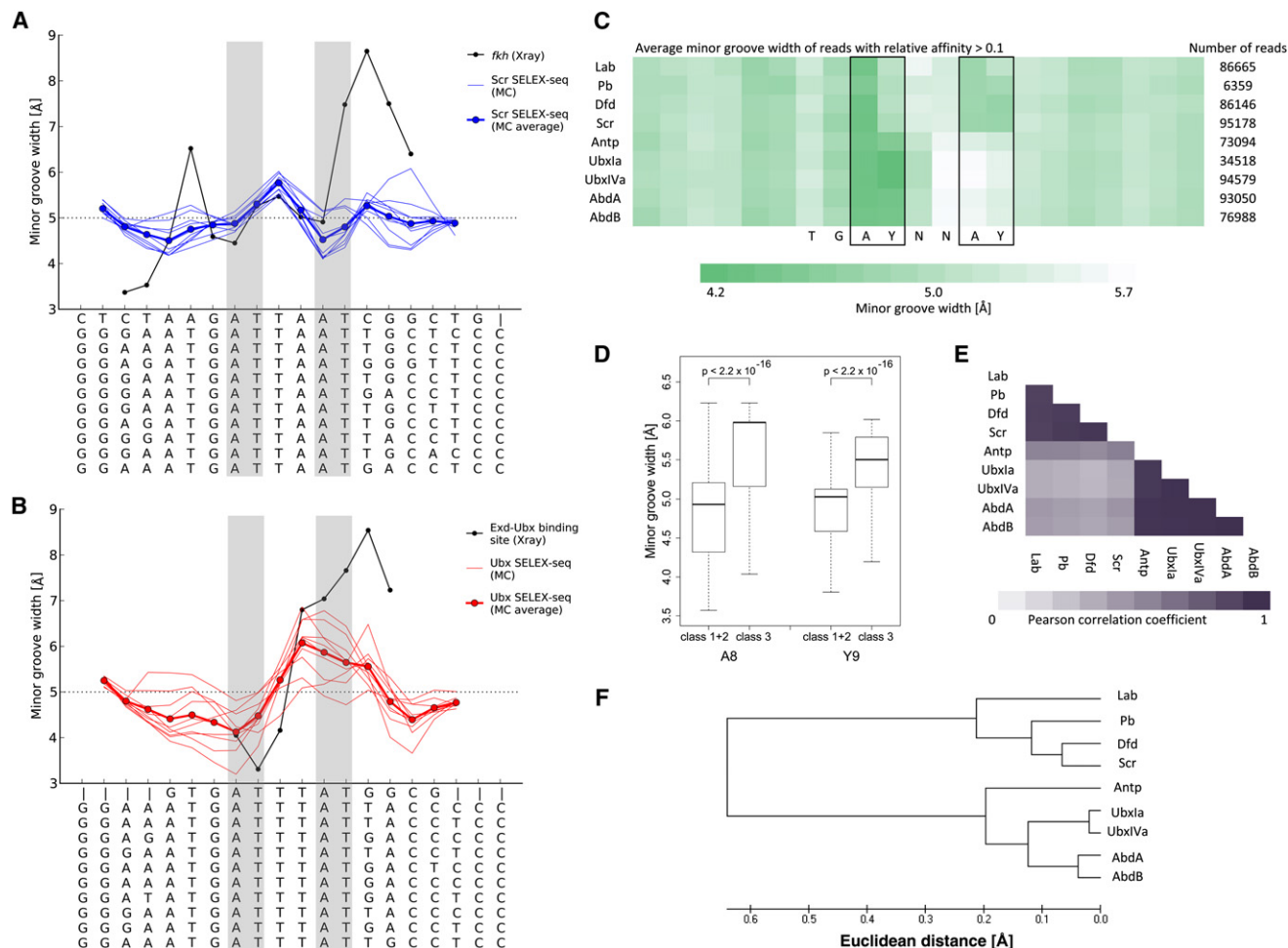


Figure 6. Predicted Minor Groove Widths of Exd-Hox Binding Sites

(A and B) MC predictions of minor groove width of selected binding sites for Exd-Scr (A) and Exd-Ubx (B). Groove widths of the DNA from crystal structures (black) of Exd-Hox-DNA ternary complexes (Joshi et al., 2007; Passner et al., 1999) are plotted with the widths predicted for the ten highest affinity binding sites (thin blue lines in [A] and thin red lines in [B]) and their average groove widths (thick blue line in [A] and thick red line in [B]). Sequences from crystal structures (top) and the ten SELEX-seq sites are below the x axis; gray shading highlights A_4T_5 and A_8T_9 .

(C) Heat map characterizing the average minor groove width of all sequences above a relative binding affinity threshold of 0.1 for each Exd-Hox heterodimer. Dark green represents narrow minor groove regions and white denotes wider minor grooves.

(D) Minor groove width values at the most distinct A_8 and Y_9 positions are compared in box plots for the data shown in panel (C) and Mann-Whitney U p-values between the two groups, class 1+2 and class 3 Hox binding sites, indicate significant differences.

(E) Average minor groove width is compared in all positions of the nTGAYNNAYnnn dodecamer for the different Exd-Hox sites using Pearson correlation. Dark purple represents high similarity while white characterizes low similarity.

(F) Dendrogram comparing minor groove shape for Exd-Hox binding sites based on Euclidean distances between average minor groove width in the six positions of the AYNNA core. See also Figures S4 and S6.

DNA Shape Contributes to Exd-Hox Dimer Preferences

We used all-atom Monte Carlo (MC) simulations (Joshi et al., 2007; Rohs et al., 2005) to predict the width of the minor grooves of the ten highest-affinity Exd-Hox binding sites identified by SELEX-seq for Scr and Ubx. The ten sequences with highest binding affinity for Exd-Scr, which all contain a blue (TGATTAAT) binding site, have a similar shape, with two narrow regions in the core (Figure 6A). This double-minimum pattern of minor groove width is similar to that seen in the crystal structure of Exd-Scr bound to the sequence *fkh250* (Joshi et al., 2007). In contrast, the predicted minor groove shapes of the ten sequences with

highest binding affinities for Exd-Ubx, which all contain a red (TGATTTAT) binding site, have a narrow minor groove in the A_4T_5 region and a relatively wide groove in the A_8T_9 region (Figure 6B). Again, this pattern mirrors that observed in an X-ray structure of Exd-Ubx bound to DNA containing the core sequence of the red motif (Passner et al., 1999; Rohs et al., 2009) and in a structure of Exd-Scr bound to a red binding site (Joshi et al., 2007).

We extended our structural analysis to additional binding sites defined in Figure 2C (Figure S4). While most sequences are predicted to have minima in the A_4Y_5 region, which likely

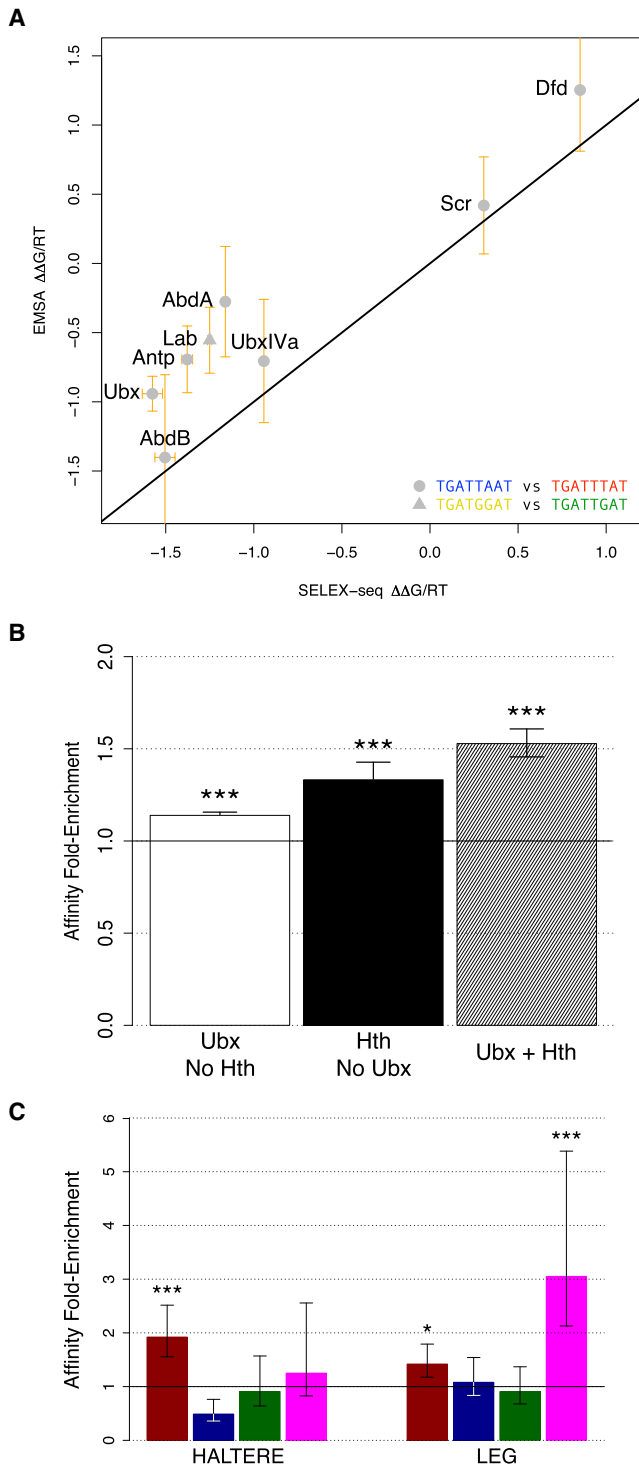


Figure 7. Relative Affinities Defined by SELEX-Seq Match Those Measured In Vitro and Correlate with Binding in Vivo

(A) Plot comparing the ratio of Kds defined by EMSA (y axis) with the ratio of relative affinities defined by SELEX-seq (x axis). Error bars on the y axis were computed using linearization and are based on the standard error of the mean over replicates for individual binding constants; those on the x axis are based on the standard error in the estimated relative affinity (see [Extended Experimental Procedures](#)). The circles represent the blue/red affinity ratios for AbdB,

accommodate the conserved Arg5 residues of both Exd and Hox, the largest variation between these binding sites is in the A_8Y_9 region. These results suggest that minor groove width is an important factor in the determination of Exd-Hox specificity. To test if this conclusion, which is based on a limited number of high-affinity sites, holds true for a large number of sequences, we employed a high-throughput approach that predicts minor groove width based on the average conformations of tetra- and pentanucleotides derived from > 1600 MC simulations ([Experimental Procedures](#)). Nearly all sequences, independent of Hox protein, had a minimum near A_4 . In contrast, binding sites preferred by class 1 and 2 Hox proteins had on average narrow minor grooves at A_8Y_9 , while those preferred by class 3 Hox proteins had on average wide minor grooves at A_8T_9 or A_8C_9 (Figure 6C). The difference in the distribution of minor groove widths at the A_8Y_9 position between class 1 plus 2 versus class 3 sequences (Figure 6D) is highly statistically significant ($p < 2.2 \times 10^{-16}$; Mann-Whitney U test). Pearson correlations between minor groove width profiles along the central 12-mer also confirm this difference in shape (Figure 6E). Clustering based on the Euclidian distance between shape profiles along the central AYNAY motif was found to be compatible with the collinear ordering of the Hox proteins from anterior to posterior (Figure 6F). This result is remarkable as it stems only from the predicted minor groove shapes of the SELEX-seq-derived binding sites.

Validation of Exd-Hox DNA Binding Site Preferences In Vitro and In Vivo

To validate these data, we compared the SELEX-seq-derived relative affinities to EMSA measurements of binding affinities (Kds) for a subset of Exd-Hox proteins, bound to the red, blue, green, or yellow motifs (Figure 7A). There is excellent agreement between the relative binding free energies derived from the two different experimental protocols (adjusted $R^2 = 0.88$). We also confirmed the preference of Exd-AbdA for a magenta (ATTAAC) binding site over a black (ATAAAT) binding site, both previously unknown Exd-Hox binding sites (Figure S5A).

A second test of the SELEX-seq results is to determine how well they predict which sites are bound by Hox proteins in vivo. Genome-wide chromatin immunoprecipitation (ChIP) has recently been used to identify Ubx binding sites in the leg and haltere imaginal discs, precursors to the ventral and dorsal third thoracic segment of the adult fly (Slattery et al., 2011). The same study also profiled Homothorax (Hth), an obligatory cofactor of Exd in vivo (Abu-Shaar et al., 1999; Rieckhof et al., 1997).

Ubxla, UbxIVa, Antp, AbdA, Scr, and Dfd. The triangle shows the yellow/green affinity ratio for Lab (adjusted $R^2 = 0.88$).

(B) Bar graphs showing the total in vitro binding affinity for Exd-Ubx (as predicted using 12-mer relative affinities derived from SELEX-seq) in genomic windows occupied in vivo by Ubx, Hth, or both (as determined using ChIP-chip), as a fold-enrichment relative to a set of control regions of the same size. Results are shown for ChIP data combined from the T3 leg and haltere. The symbols above each bar denote the statistical significance level (***) $p < 0.001$, * $p < 0.05$. Error bars correspond to standard errors, computed based on a thousand samples from the control distribution.

(C) Same as (B), but separated by core motif color and tissue (haltere on the left, T3 leg on the right). See also Figure S5.

Thus, overlapping ChIP peaks for Ubx and Hth are indicative of binding by Exd-Ubx heterodimers. Using the 12-mer tables derived from our SELEX-seq data, we summed the predicted relative binding affinity for Exd-Ubx in a sliding 12-mer window across all Ubx-bound genomic regions. Comparison of this sum to a null distribution obtained using random sampling of nearby control regions of equal size showed statistically significant enrichment (Figure 7B). As expected, the enrichment was largest when we restricted the analysis to regions bound by both Ubx and Hth (Figure 7B). To validate the relative preference of Exd-Ubx for the various core motifs, we repeated the above analysis, but summing the predicted affinity only over 12-mer windows matching a particular motif color. Comparison to the null distribution confirmed enrichment for red and magenta sites, but not green or blue, consistent with the SELEX-seq data (Figure 7C, see also Figure S5B). Moreover, the data indicate that Exd-Ubx prefers red binding sites in the haltere and magenta binding sites in the leg, suggesting that these motifs may be used in a tissue-specific manner in vivo (Figure 7C).

DISCUSSION

We have demonstrated that, compared to their monomeric DNA binding specificities, individual members of the Hox protein family acquire novel DNA recognition properties when they bind together with the cofactor Exd. As such, these results provide a precedent for how interactions between DNA binding proteins can result in emergent recognition properties that are not exhibited by either factor on their own. Based on these findings, we propose that other combinations of DNA binding proteins may use “latent specificity” strategies to achieve specificity in vivo. Thus, while there are likely to be many factors that influence binding site occupancy and transcription factor activities in vivo, our observations help bridge the gap between the degenerate specificities of monomeric transcription factors observed in vitro with the more restricted functional binding that is typically observed in vivo.

In vitro selection coupled with high-throughput sequencing has been used previously to estimate k-mer based affinity tables for monomeric transcription factors (Jolma et al., 2010; Zhao et al., 2009; Zykovich et al., 2009). However, the SELEX-seq methodology presented here has made it possible to quantify DNA recognition by transcription factor complexes at full resolution. The use of EMSA allowed us to focus on a specific and cooperative heterodimeric complex. Statistical modeling of the composition of the initial pool, together with integration of multiple early rounds of selection, allowed quantification of relative DNA binding affinities for all specifically bound 12-base-pair sequences over almost two orders of magnitude. Together, these methods provide an ideal framework for analyzing the DNA binding preferences for transcription factor complexes.

A Single Cofactor Can Reveal Latent DNA Binding Specificities that Distinguish Members of the Same Transcription Factor Family

As monomers, the eight Hox proteins in *Drosophila* recognize an overlapping set of AT-rich hexameric binding sites (Mann et al., 2009; Noyes et al., 2008). In the presence of Exd, however, we

find that Hox DNA binding preferences become more focused and specific. These findings raise an important question: how can the same cofactor elicit unique specificities for eight closely related homeodomain proteins? We propose that the additional specificity information that is used to distinguish Exd-Hox binding preferences comes from the Hox protein, but that this information cannot be used effectively without Exd. In other words, Exd unlocks latent specificities that are present within the Hox protein sequences. It is plausible that other protein families use an analogous mechanism to fine tune their DNA binding specificities. For example, Runt domain proteins bind DNA with a higher degree of specificity when partnered with the cofactor CBF β (core-binding factor) and different combinations of bHLH proteins appear to prefer different E-box sequences (Bartfeld et al., 2002; Grove et al., 2009; Tahirov et al., 2001). We further speculate that novel specificities may emerge as a consequence of the assembly of higher order multi-protein-DNA complexes.

How might this work in molecular terms? For Hox proteins, one source of latent specificity information is thought to be in the N-terminal arms of their homeodomains and neighboring linker sequences. By binding the “YPWM” motif, which is located N-terminal to Hox homeodomains (Figure 2C), Exd limits the structural freedom of this portion of the Hox protein. For Scr, the YPWM-Exd interaction positions this region of the Hox protein so that it can bind to the minor groove, primarily via three basic residues: two Arginines (Arg3 and Arg5 of the homeodomain) and a Histidine (His-12) (Joshi et al., 2007). Importantly, several residues in Scr’s N-terminal arm and linker region are conserved in a paralog-specific manner and are important for executing Scr-specific functions in vivo (Joshi et al., 2007). Some of these residues correlate with the binding specificities identified here. For example, both class 2 Hox proteins (Dfd and Scr), but none of the other Hox proteins, have His at position –12 (numbering is from the start of the homeodomain; Figure S6A). Further, only class 2 Hox proteins have the N-terminal arm motif “RQR” (where the first Arg is Arg3; Figure S6A). Although most other Hox proteins have an Arg at position 3, the adjacent Gln is unique to class 2 proteins and is required for optimal binding, perhaps by favoring a conformation in which both Arg3 and Arg5 can insert into the minor groove (Joshi et al., 2007). Based on these correlations, we suggest that the RQR motif contributes to the preference that class 2 proteins exhibit in our SELEX-seq experiments (Figure S6C). Additional correlations between Hox protein sequences and SELEX binding site preferences are also apparent (Figure S6). For example, all class 3b Hox proteins (Ubx, AbdA, and AbdB) have an Arg at position 2 of the homeodomain. In a crystal structure of the vertebrate AbdB ortholog HoxA9 bound to DNA in complex with Pbx, this Arg makes multiple water-mediated hydrogen bonds in the minor groove of a magenta binding site (Figure S6E) (LaRonde-LeBlanc and Wolberger, 2003; Mann et al., 2009). Together, these observations suggest that seemingly small differences in protein sequence between Hox proteins are exploited by Exd to help achieve DNA binding specificity.

Despite their importance, Hox homeodomain and linker sequences are unlikely to account for all of the differences we

observe between Exd-Hox complexes. One reason is that the proteins used in all of the existing crystal structures are primarily limited to the DNA binding domains, while the proteins used in the SELEX-seq experiments are significantly longer and in many cases, close to full-length. In vitro, the protein fragments used in the crystal structures bind to their binding sites with significantly less cooperativity than full-length proteins, suggesting that additional interactions are likely to exist in the native complexes (unpublished data). Additional structural studies using full-length proteins and alternative binding sites will be needed to fully understand the specificities revealed here.

The Role of DNA Shape in Hox DNA Recognition

Several lines of evidence suggest that discrimination of specific DNA sequences by proteins depends in part on the recognition of sequence-dependent differences in DNA structure, such as groove width (Rohs et al., 2010). In the present work, we find that all preferred binding sites, regardless of Exd-Hox preference, are predicted to have narrow minor grooves at TGAY (positions 2 to 5). In all of the existing crystal structures, Arg5 of both Exd and Hox are either bound to or located near to this narrow minor groove region, likely mediated through electrostatic interactions (Rohs et al., 2009).

In contrast to this shared feature, minor groove topography varies in the Hox portion of these binding sites. Most notably, class 1 and 2 Hox proteins select binding sites that have an additional minor groove minimum close to the AY of the Hox half site, NNAY, whereas class 3 Hox proteins prefer a wider minor groove in this region. In several cases the binding sites preferred by a particular Exd-Hox complex have similar DNA shapes despite having different sequences, in agreement with the observation that DNA shape is often more conserved than DNA sequence (Parker et al., 2009). That minor groove shape may play an important role in Exd-Hox binding preferences is further underscored by our observation that this parameter was sufficient to partition the preferred binding sites of the three classes of Hox proteins, irrespective of the primary sequence.

It is interesting that most of the sequence variation contributing to Hox preference is located at positions 6 and 7 (Figure 2C). Remarkably, the base pair at position 7 makes no protein contacts in any of the known crystal structures, while position 6 makes only a small number of contacts that do not appear to be specific. How is it possible that a single nucleotide position that makes no contacts can play such an important role in specificity? We suggest that the effect is due to the location of a TpR step (R = A or G), which tends to widen the minor groove (Joshi et al., 2007). There is a TpR step at positions 6 and 7 in most class 1 and 2 sites that should widen the groove in the middle of the binding site, allowing Arg3 and Arg5 to bind to the two minima on either side. In contrast, the TpA step in most class 3 sites at positions 7 and 8 may block Arg3 from stably inserting into the groove. Note that the shift of the TpR step by one nucleotide in the 3' direction is the main source of variability at position 7 since there is a purine at this position in class 1 and class 2 sites and T in class 3 sites.

Constraints on the Evolution of Exd-Hox Binding Preferences

When the first complex of Hox genes was discovered in *Drosophila*, it was realized that the order of Hox genes along the chromosome was collinear with their corresponding functional domains along the AP axis of the adult fly (Lewis, 1978). Collinearity was later extended to Hox expression patterns along the AP axis during fly and vertebrate embryogenesis (McGinnis and Krumlauf, 1992). Here, we extend this rule further by showing that differences in the DNA binding specificities of Exd-Hox complexes, as well as the minor groove topographies of their preferred binding sites, are also collinear with these other Hox gene characteristics. Collinearity of DNA binding preferences likely reflects the way in which the Hox genes duplicated and gradually diverged during evolution (Hueber et al., 2010; McGinnis and Krumlauf, 1992).

When presented with all possible 16-mers, the preferred binding sites for each of the nine Exd-Hox complexes characterized here share the structure WRAYNNAY. This binding site matches nearly all of the known in vivo binding sites for Exd-Hox or Pbx-Hox complexes (Mann et al., 2009). Thus, it appears that for HM-Exd-Hox complexes, alternative modes of binding are not used by these factors. These observations suggest that the biophysical properties of these proteins have constrained the evolution of Exd-Hox-DNA interactions. Moreover, the preferred binding sites identified by SELEX-seq are present in bona fide in vivo binding sites that have been characterized by more traditional methods: for example, Exd-Scr regulates its target *forkhead (fkh)* via a blue binding site and Exd-Lab autoregulates *labial* via a yellow binding site (Ryoo and Mann, 1999; Ryoo et al., 1999). We also found that on a genome-wide level, regions bound in vivo by Exd-Ubx are specifically enriched in red and magenta binding sites. Although chromatin structure and interactions with other proteins in vivo no doubt also influence Hox binding and activity, these findings suggest that the Exd-Hox binding site signatures identified here will be important for deciphering the sequence determinants that guide the binding, and eventually the function, of these proteins in vivo.

EXPERIMENTAL PROCEDURES

A full description of the methods is in the [Supplemental Information](#).

SELEX

Hox and HM-Exd purification conditions, expression constructs, and EMSA conditions have been described previously (Joshi et al., 2010; Noro et al., 2006; Ryoo and Mann, 1999). AbdB (residue 224 to the C terminus) and Pbx (residues 126–306) were cloned in pET14b (Novagen) and pQE9 (QIAGEN), respectively. For SELEX EMSA lanes, binding reactions were performed with 200 nM double-stranded SELEX library, 67 nM Hox, and 33 nM HM-Exd in a final volume of 30 μ l. Parallel DNA binding reactions using ³²P labeled probes containing known Hox-Exd composite sites were used to track the mobility of Hox+HM-Exd+DNA complexes. Regions corresponding to the cooperative complex were cut out and eluted. The eluted DNA was purified, concentrated, and amplified by PCR. The PCR products were then purified and divided for the next round of SELEX or Illumina sequencing. Subsequent rounds of selection followed the same structure as the first round of SELEX (Figure 1 and Extended Experimental Procedures).

Inferring Relative Affinities

A 5th-order Markov model was constructed using the sequences in R0 and used to predict the expected number of occurrences of each 12-mer in R0. The fold enrichment from R0 to R1 for a particular 12-mer was computed as the ratio of the actual count in R1 and the predicted count in R0. A first estimate of relative affinity for each 12-mer was obtained by normalizing by the highest fold-enrichment. A second, independent, estimate was proportional to the square root of the fold-enrichment from R2 and R0 (or the cubic root of that between R3 and R0). To correct for non-linear bias in the later round, local regression of the R2vsR0 (or R3vsR0) affinities on the R1vsR0 affinities was performed, and used to transform the former to a final estimate of relative 12-mer affinities.

DNA Shape Analysis

DNA structures were predicted using all-atom Monte Carlo (MC) simulations (Rohs et al., 2005; Joshi et al., 2007) without the protein present. For the high-throughput shape analysis, a total of 1,658 trajectories from independent MC simulations were used to build a database of shape predictions, based on the conformation of all tetra- and pentanucleotides. All SELEX-seq reads with a relative affinity above 0.1 were aligned based on the TGAYNNAY motif (excluding reads which had more than one motif) and the average minor groove width in each position was calculated. To compare the shape of Exd-Hox sites, we calculated box plots for minor groove width in the most distinct positions A8 and Y9 for class 1+2 versus class 3 binding sites and compared the average width in all positions of the 12-mer nTGAYNNAYnnn using Pearson correlation. The width values at the six positions of the AYNNA core motif were used to calculate a Euclidean distance tree that relates the shapes selected by all Exd-Hox dimers.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, four tables, and six figures and can be found with this article online at doi:10.1016/j.cell.2011.10.053.

ACKNOWLEDGMENTS

We thank the members of the Bussemaker, Honig, Mann, and Rohs labs for comments and feedback during the course of these studies; B. Snyder for programming support; W. Shin and C. Hawkins for early contributions to the analysis of SELEX-seq data; A. Boyanov for sequencing support; and K. Lelli for constructs. This work was supported by NIH grants U54CA121852, R01GM054510, R01HG003008, R01GM030518, and P50GM071508, a John Simon Guggenheim Foundation Fellowship, Columbia University's RISE program, USC start-up funds, the USC-Technion Visiting Fellows Program, and an Andrew Viterbi Fellowship. Author contributions: M.S. designed and executed the SELEX experiments and contributed to the analysis of the SELEX data. T.R. designed and executed the analysis of the SELEX-seq data. P.L. carried out and analyzed Monte Carlo simulations. N.A. carried out later rounds of SELEX and EMSA-based validation experiments. P.G.-A. analyzed Ubx ChIP data. I.D. executed high-throughput DNA shape analysis. T.Z. developed high-throughput DNA shape prediction method. R.R. and B.H. supervised and designed DNA shape analyses. H.J.B. supervised and designed the analysis of the SELEX-seq data. R.S.M. supervised and designed SELEX experiments; contributed to the analysis of the SELEX data and shape analyses.

Received: June 9, 2011

Revised: August 19, 2011

Accepted: October 6, 2011

Published: December 8, 2011

REFERENCES

Abu-Shaar, M., Ryoo, H.D., and Mann, R.S. (1999). Control of the nuclear localization of Extradenticle by competing nuclear import and export signals. *Genes Dev.* 13, 935–945.

Alexander, T., Nolte, C., and Krumlauf, R. (2009). Hox genes and segmentation of the hindbrain and axial skeleton. *Annu. Rev. Cell Dev. Biol.* 25, 431–456.

Bartfeld, D., Shimon, L., Couture, G.C., Rabinovich, D., Frolow, F., Levanon, D., Groner, Y., and Shakked, Z. (2002). DNA recognition by the RUNX1 transcription factor is mediated by an allosteric transition in the RUNT domain and by DNA bending. *Structure* 10, 1395–1407.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* 133, 1266–1276.

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G.J., Parker, M.H., MacQuarrie, K.L., Davison, J., Morgan, M.T., Ruzzo, W.L., et al. (2010). Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Dev. Cell* 18, 662–674.

Chan, S.K., Jaffe, L., Capovilla, M., Botas, J., and Mann, R.S. (1994). The DNA binding specificity of Ultrabithorax is modulated by cooperative interactions with extradenticle, another homeoprotein. *Cell* 78, 603–615.

Chang, C.P., Brocchieri, L., Shen, W.F., Largman, C., and Cleary, M.L. (1996). Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. *Mol. Cell. Biol.* 16, 1734–1745.

Conlon, F.L., Fairclough, L., Price, B.M., Casey, E.S., and Smith, J.C. (2001). Determinants of T box protein specificity. *Development* 128, 3749–3758.

Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. *Nat. Genet.* 42, 255–259.

Gebelein, B., McKay, D.J., and Mann, R.S. (2004). Direct integration of Hox and segmentation gene inputs during Drosophila development. *Nature* 431, 653–659.

Gehring, W.J., Qian, Y.Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A.F., Resendez-Perez, D., Affolter, M., Otting, G., and Wuthrich, K. (1994). Homeodomain-DNA recognition. *Cell* 78, 211–223.

Grove, C.A., De Masi, F., Barrasa, M.I., Newburger, D.E., Alkema, M.J., Bulyk, M.L., and Walhout, A.J. (2009). A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* 138, 314–327.

Guertin, M.J., and Lis, J.T. (2010). Chromatin landscape dictates HSF binding to target DNA elements. *PLoS Genet.* 6, e1001114.

Hanes, S.D., and Brent, R. (1989). DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. *Cell* 57, 1275–1283.

Hueber, S.D., Weiller, G.F., Djordjevic, M.A., and Frickey, T. (2010). Improving Hox protein classification across the major model organisms. *PLoS ONE* 5, e10820.

Johnson, A.D. (1995). Molecular mechanisms of cell-type determination in budding yeast. *Curr. Opin. Genet. Dev.* 5, 552–558.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpaa, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* 20, 861–873.

Jones, S. (2004). An overview of the basic helix-loop-helix proteins. *Genome Biol.* 5, 226.

Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B., and Mann, R.S. (2007). Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* 131, 530–543.

Joshi, R., Sun, L., and Mann, R. (2010). Dissecting the functional specificities of two Hox proteins. *Genes Dev.* 24, 1533–1545.

Kaplan, T., Li, X.Y., Sabo, P.J., Thomas, S., Stamatoyannopoulos, J.A., Biggin, M.D., and Eisen, M.B. (2011). Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLoS Genet.* 7, e1001290.

- LaRonde-LeBlanc, N.A., and Wolberger, C. (2003). Structure of HoxA9 and Pbx1 bound to DNA: Hox hexapeptide and DNA recognition anterior to posterior. *Genes Dev.* *17*, 2060–2072.
- Lewis, E.B. (1978). A gene complex controlling segmentation in *Drosophila*. *Nature* *276*, 565–570.
- Li, X., and McGinnis, W. (1999). Activity regulation of Hox proteins, a mechanism for altering functional specificity in development and evolution. *Proc. Natl. Acad. Sci. USA* *96*, 6802–6807.
- Lu, Q., and Kamps, M.P. (1997). Heterodimerization of Hox proteins with Pbx1 and oncoprotein E2a-Pbx1 generates unique DNA-binding specificities at nucleotides predicted to contact the N-terminal arm of the Hox homeodomain—demonstration of Hox-dependent targeting of E2a-Pbx1 in vivo. *Oncogene* *14*, 75–83.
- Lutz, B., Lu, H.C., Eichele, G., Miller, D., and Kaufman, T.C. (1996). Rescue of *Drosophila* labial null mutant by the chicken ortholog Hoxb-1 demonstrates that the function of Hox genes is phylogenetically conserved. *Genes Dev.* *10*, 176–184.
- Mann, R.S. (1995). The specificity of homeotic gene function. *Bioessays* *17*, 855–863.
- Mann, R.S., and Chan, S.K. (1996). Extra specificity from extradenticle: the partnership between HOX and PBX/EXD homeodomain proteins. *Trends Genet.* *12*, 258–262.
- Mann, R.S., Lelli, K.M., and Joshi, R. (2009). Hox specificity unique roles for cofactors and collaborators. *Curr. Top. Dev. Biol.* *88*, 63–101.
- McGinnis, N., Kuziora, M.A., and McGinnis, W. (1990). Human Hox-4.2 and *Drosophila* deformed encode similar regulatory specificities in *Drosophila* embryos and larvae. *Cell* *63*, 969–976.
- McGinnis, W., and Krumlauf, R. (1992). Homeobox genes and axial patterning. *Cell* *68*, 283–302.
- Naiche, L.A., Harrelson, Z., Kelly, R.G., and Papaioannou, V.E. (2005). T-box genes in vertebrate development. *Annu. Rev. Genet.* *39*, 219–239.
- Noro, B., Culi, J., McKay, D.J., Zhang, W., and Mann, R.S. (2006). Distinct functions of homeodomain-containing and homeodomain-less isoforms encoded by homothorax. *Genes Dev.* *20*, 1636–1650.
- Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and Wolfe, S.A. (2008). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* *133*, 1277–1289.
- Parker, S.C., Hansen, L., Abaan, H.O., Tullius, T.D., and Margulies, E.H. (2009). Local DNA topography correlates with functional noncoding regions of the human genome. *Science* *324*, 389–392.
- Passner, J.M., Ryoo, H.D., Shen, L., Mann, R.S., and Aggarwal, A.K. (1999). Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. *Nature* *397*, 714–719.
- Pearson, J.C., Lemons, D., and McGinnis, W. (2005). Modulating Hox gene functions during animal body patterning. *Nat. Rev. Genet.* *6*, 893–904.
- Piper, D.E., Batchelor, A.H., Chang, C.P., Cleary, M.L., and Wolberger, C. (1999). Structure of a HoxB1-Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation. *Cell* *96*, 587–597.
- Rieckhof, G.E., Casares, F., Ryoo, H.D., Abu-Shaar, M., and Mann, R.S. (1997). Nuclear translocation of extradenticle requires homothorax, which encodes an extradenticle-related homeodomain protein. *Cell* *91*, 171–183.
- Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* *79*, 233–269.
- Rohs, R., Sklenar, H., and Shakked, Z. (2005). Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure* *13*, 1499–1509.
- Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. *Nature* *461*, 1248–1253.
- Ryoo, H.D., and Mann, R.S. (1999). The control of trunk Hox specificity and activity by Extradenticle. *Genes Dev.* *13*, 1704–1716.
- Ryoo, H.D., Marty, T., Casares, F., Affolter, M., and Mann, R.S. (1999). Regulation of Hox target genes by a DNA bound Homothorax/Hox/Extradenticle complex. *Development* *126*, 5137–5148.
- Slattery, M., Ma, L., Negre, N., White, K.P., and Mann, R.S. (2011). Genome-wide tissue-specific occupancy of the hox protein ultrabithorax and hox cofactor homothorax in *Drosophila*. *PLoS ONE* *6*, e14686.
- Taghli-Lamallem, O., Gallet, A., Leroy, F., Malapert, P., Vola, C., Kerridge, S., and Fasano, L. (2007). Direct interaction between Teashirt and Sex combs reduced proteins, via Tsh's acidic domain, is essential for specifying the identity of the prothorax in *Drosophila*. *Dev. Biol.* *307*, 142–151.
- Tahirov, T.H., Inoue-Bungo, T., Morii, H., Fujikawa, A., Sasaki, M., Kimura, K., Shiina, M., Sato, K., Kumasaka, T., Yamamoto, M., et al. (2001). Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta. *Cell* *104*, 755–767.
- Treisman, J., Gonczy, P., Vashishtha, M., Harris, E., and Desplan, C. (1989). A single amino acid can determine the DNA binding specificity of homeodomain proteins. *Cell* *59*, 553–562.
- Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* *249*, 505–510.
- Wunderlich, Z., and Mirny, L.A. (2009). Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet.* *25*, 434–440.
- Zhao, J.J., Lazzarini, R.A., and Pick, L. (1993). The mouse Hox-1.3 gene is functionally equivalent to the *Drosophila* Sex combs reduced gene. *Genes Dev.* *7*, 343–354.
- Zhao, Y., Granas, D., and Stormo, G.D. (2009). Inferring binding energies from selected binding sites. *PLoS Comput. Biol.* *5*, e1000590.
- Zykovich, A., Korf, I., and Segal, D.J. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic Acids Res.* *37*, e151.