

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

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

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

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

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

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

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

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

<sup>7</sup>These authors contributed equally to this work

<sup>8</sup>Present address: Institute for Genomics and Systems Biology, University of Chicago, 900 East 57<sup>th</sup> 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

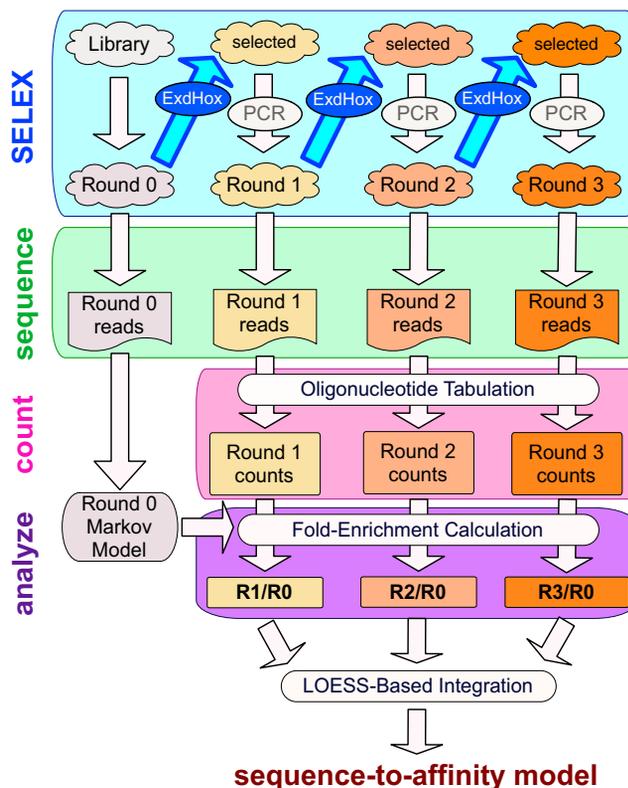
Gene regulatory information is encoded in genomic DNA sequences and interpreted by transcription factors that bind to specific sequences. Although the in vitro binding properties of transcription factors have been studied for many years, it has proven notoriously difficult to predict in vivo genomic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

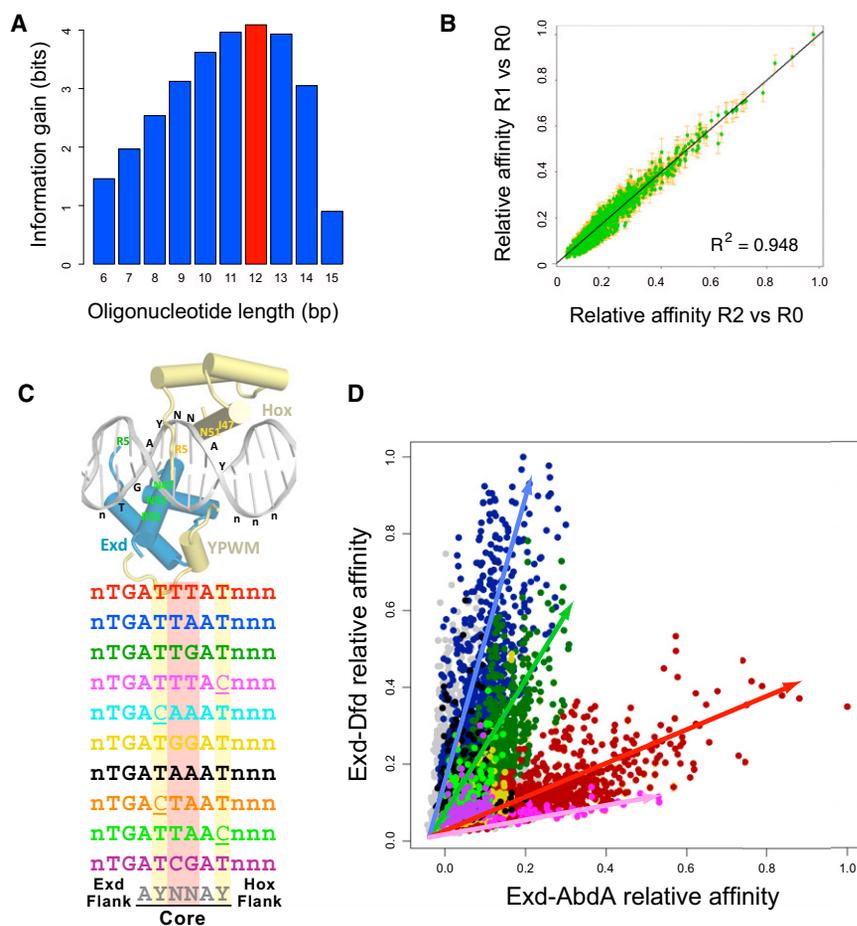
To determine the number of consecutive base pairs that must be specified to fully capture Exd-Hox binding specificities we

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

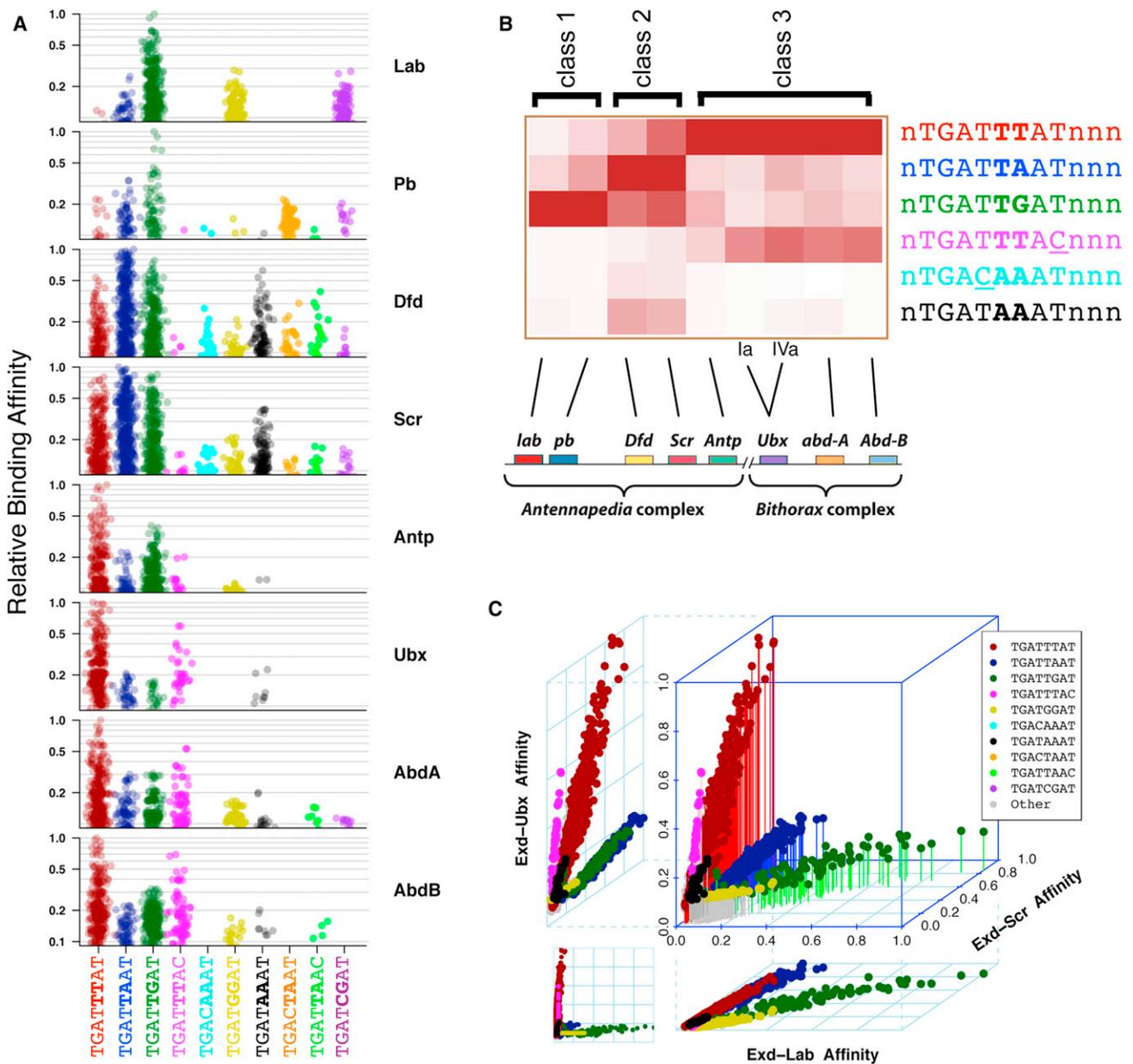
The highest-affinity 12-mer for each Exd-Hox dimer (cf. [Table S2](#)) has the form ATGATTDATNNN (where D = G, A, or T). As seen in X-ray structures, the first and second underlined A:T base pairs are contacted by Asn51 of the  $\alpha$ 3 recognition helices of the Exd and Hox homeodomains in the major 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](#)).

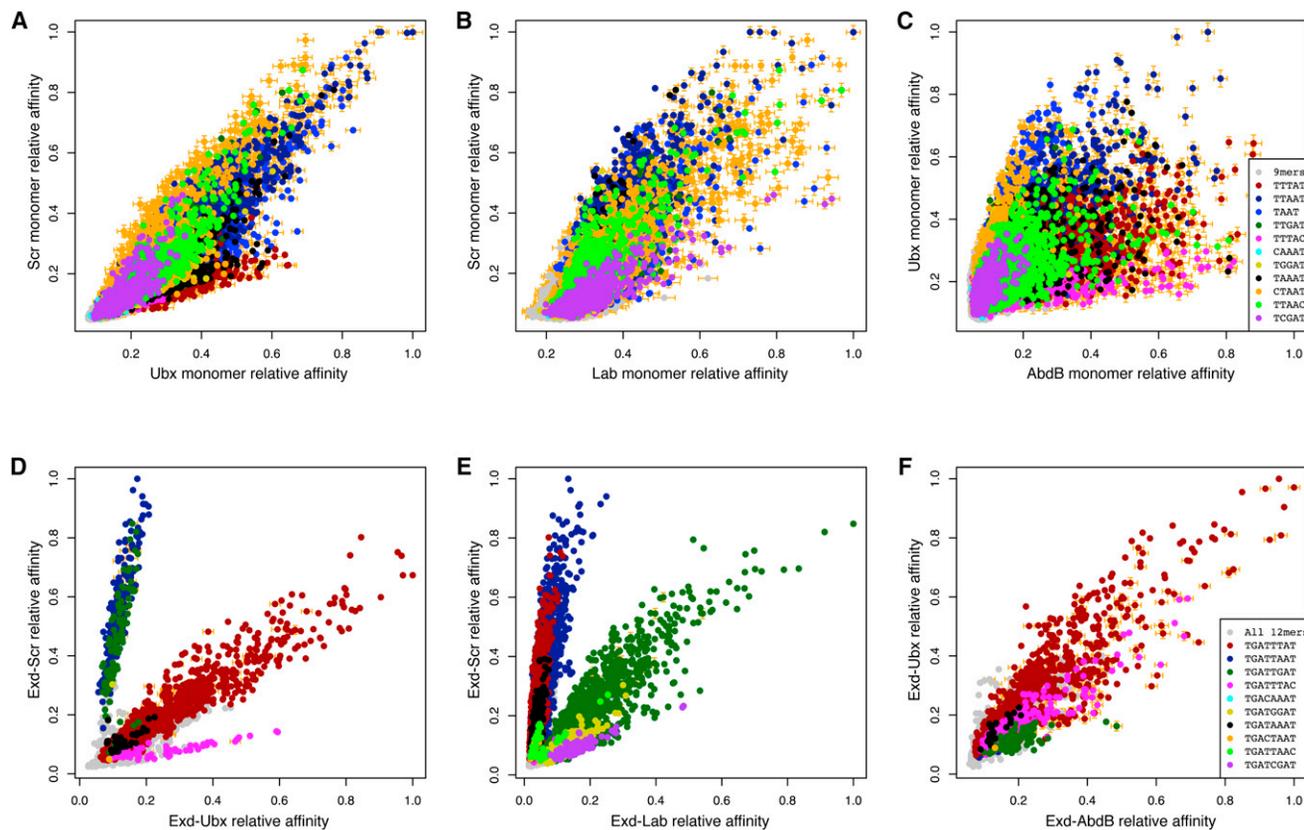


**Figure 3. Exd-Hox Heterodimers Can Be Distinguished Based on Their DNA Specificity Fingerprints**

(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](#).

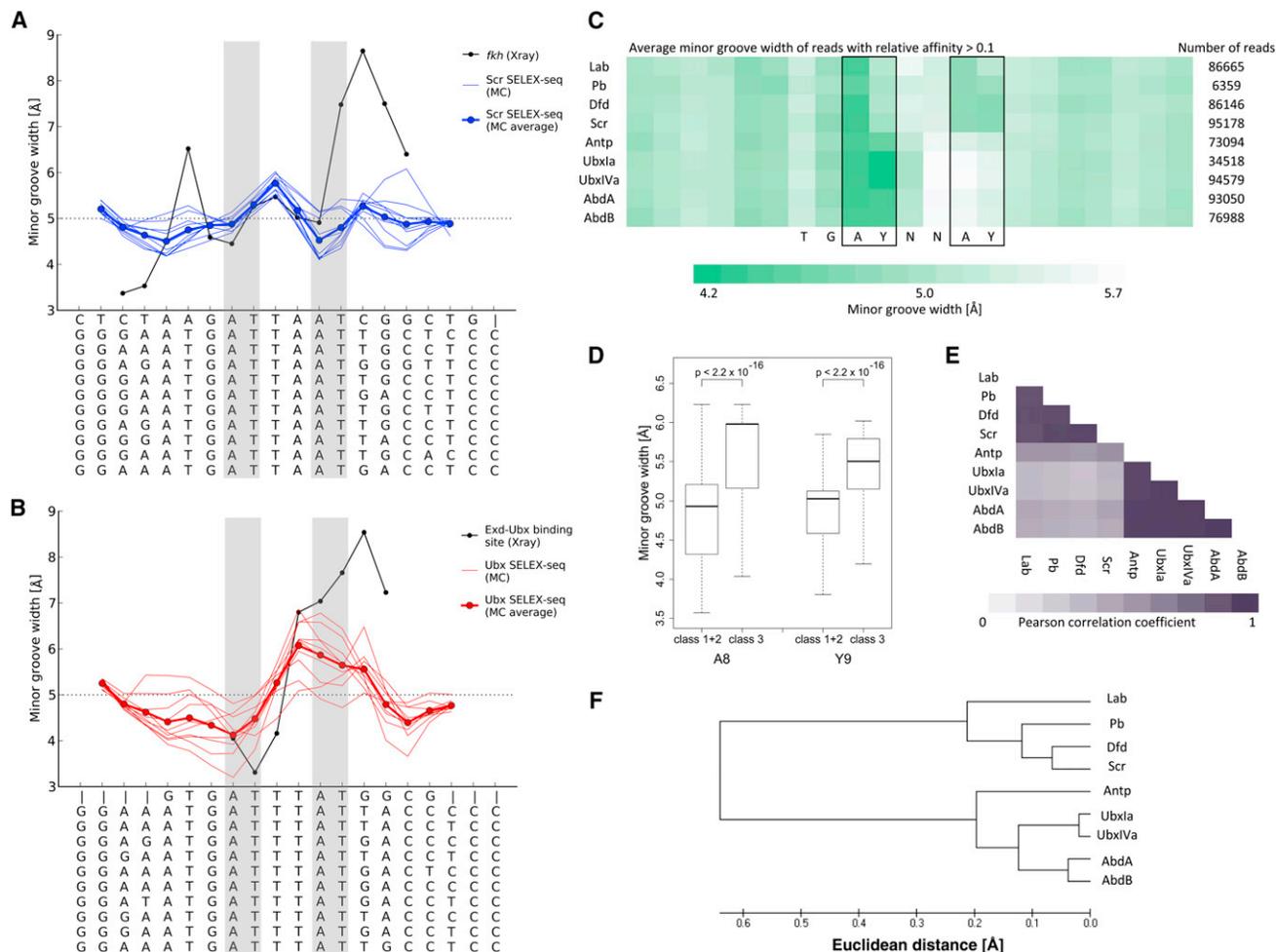
### Unique Hox DNA Binding Preferences Are Revealed upon Heterodimerization with Exd

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 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.

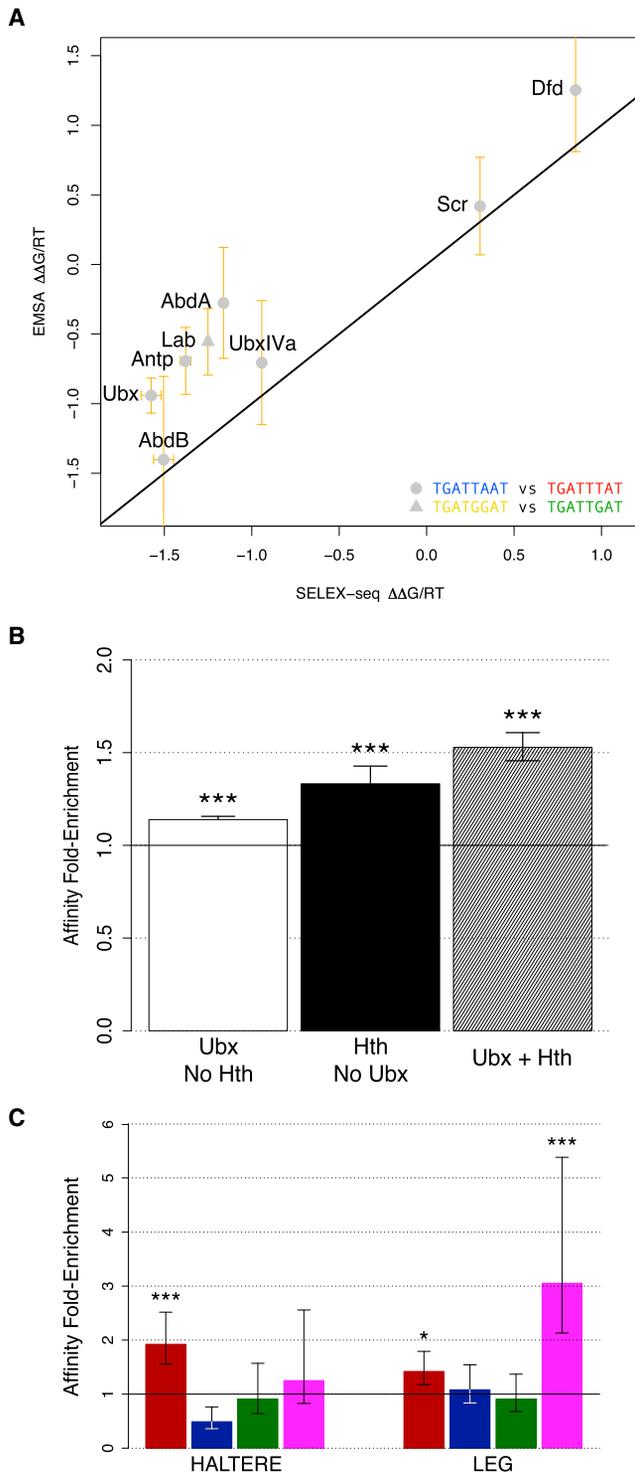
(F) Dendrogram comparing minor groove shape for Exd-Hox binding sites based on Euclidean distances between average minor groove width in the six positions of the AYNAY core. See also Figures S4 and S6.

### 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<sub>8</sub>Y<sub>9</sub> 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<sub>4</sub>. In contrast, binding sites preferred by class 1 and 2 Hox proteins had on average narrow minor grooves at A<sub>8</sub>Y<sub>9</sub>, while those preferred by class 3 Hox proteins had on average wide minor grooves at A<sub>8</sub>T<sub>9</sub> or A<sub>8</sub>C<sub>9</sub> ([Figure 6C](#)). The difference in the distribution of minor groove widths at the A<sub>8</sub>Y<sub>9</sub> 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 $\beta$  (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  $\mu$ l. Parallel DNA binding reactions using <sup>32</sup>P 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 5<sup>th</sup>-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.

## EXTENDED EXPERIMENTAL PROCEDURES

### Oligonucleotides

All oligonucleotides referenced in the [Experimental Procedures](#) are listed in [Table S1](#).

### Protein Purification and EMSAs

All proteins were purified from BL21 bacteria as His-tagged fusions using Ni-chromatography as described previously (Gebelein et al., 2002) ([Figure S2A](#)). His-tagged, full length Exd was copurified with the HM domain of Hth (Noro et al., 2006). Lab (Chan et al., 1996), Dfd and Scr (Joshi et al., 2010), Antp (Jaffe et al., 1997), Ubx1a (Ryoo and Mann, 1999), Ubx1Va (Gebelein et al., 2002), and AbdA (Ryoo and Mann, 1999) have been described. AbdB (residue 224 to the carboxyl terminus) was cloned in frame with the His tag of pET14b (Novagen), and Pb (residues 126–306) was cloned in frame with the His tag of pQE9 (QIAGEN). AbdB and Pb cDNAs were generous gifts from Bill McGinnis and David Cribbs, respectively. Of all the Hox proteins, Pb was the most difficult to purify and showed the least amount of cooperative binding with HM-Exd, which likely accounts for the lower frequency of Exd-Hox binding sites in the Exd-Pb selected oligos. Electrophoretic mobility shift assays (EMSAs) were performed as described (Gebelein et al., 2002). For SELEX EMSA lanes, binding reactions were performed with 200 nM double-stranded SELEX library (described below), 67 nM Hox, and 33 nM HM-Exd in a final volume of 30  $\mu$ l. Parallel DNA binding reactions using  $^{32}$ P labeled probes containing known Hox-Exd composite sites were used to track the mobility of Hox+HM-Exd+DNA complexes (described below). For Kd measurements, increasing amounts of the Hox protein (from 5–800 nM) were added to a reaction mix with 80 nM HM-Exd, and the data were analyzed as described (Joshi et al., 2010).

### SELEX

The 73 bp oligonucleotides “SELEX 16mer Multiplex 1” and “SELEX 16mer Multiplex 2” including 16 random nucleotides, two PCR primer sequences and three bases of barcode sequence for multiplexing were synthesized by Integrated DNA Technologies using the hand-mix option for the randomized region. The corresponding double-stranded random libraries were generated by a Klenow primer extension reaction with the 73 bp oligonucleotide templates and the reverse primer “SELEX SR 1” followed by MinElute purification (QIAGEN). The invariable PCR primer sequences were designed to allow amplification of the in vitro selected library with primers “SELEX SR 0” and “SELEX SR 1,” and the barcode sequences were included to permit multiplexed Illumina sequencing (Lefrancois et al., 2009).

The binding reaction for the first round of SELEX was performed as described above (200 nM SELEX library, 67 nM Hox, 33 nM HM-Exd in a 30  $\mu$ l reaction), with parallel reactions containing radiolabeled probe to monitor the mobility of Hox+HM-Exd+DNA complexes. We also carried out SELEX-seq with only HM-Exd (no Hox), which confirmed the identity of Exd-Exd dimer sites selected in some of the Exd-Hox selections ([Figure S3](#)). The radiolabeled probes were the same size as the SELEX library and contained scrambled adaptor sequences so they would not contaminate the SELEX library during amplification. EMSA gels were dried, imaged on a phosphorimager (GE Healthcare), and regions corresponding to the cooperative complex were cut out and eluted overnight (37°C) in elution buffer (0.5 M NHOAc, 1 mM EDTA, 0.1% SDS). The eluted DNA was purified and concentrated by phenol:chloroform extraction and ethanol precipitation. Half of the eluted DNA (10  $\mu$ l) was then amplified by PCR with the primers “SELEX SR 0” and “SELEX SR 1.” For PCR, the 10  $\mu$ l of eluted DNA was split equally among five, 50  $\mu$ l reactions (0.4  $\mu$ M each primer, 0.2 mM each dNTP, 2.5 units Taq polymerase). The PCR products were then purified and 6 pmol was used for the next round of SELEX; the remainder of the purified PCR product was saved for Illumina sequencing (discussed below). Subsequent rounds of selection followed the same structure as the first round of SELEX (see [Figure 1](#)).

To prepare libraries for Illumina sequencing the amplified PCR products from each round of SELEX, and the unamplified double-stranded SELEX libraries (“SELEX 16mer Multiplex 1” and “SELEX 16mer Multiplex 2”; also called R0 in the Results section), were subjected to limited cycle PCR. This PCR step was necessary for the addition of 23 bp to the 5' end of each library (when treating the oligos described in [Table S1](#) as the plus strand), which was necessary to make the libraries compatible with an Illumina flow cell. For this limited cycle PCR, 95 ng of library DNA was split equally among five, 50  $\mu$ l reactions with 0.8  $\mu$ M each primer, and 0.3 mM each dNTP. PCR was performed with Phusion DNA polymerase (New England Biolabs). Sequencing-compatible library (96 bp) was separated from remaining library by acrylamide gel electrophoresis. The 96 bp band was cut from the gel and eluted in 1x NEB Buffer 2 (2 hr at room temperature), followed by ethanol precipitation. Purified DNA was sequenced on an Illumina GAIIx sequencer according to the Illumina protocol for small RNA cluster generation and 36 cycle sequencing. In most cases a single lane of a flow cell contained one Multiplex 1 library and one Multiplex 2 library.

### Modeling Biases in the Initial DNA Pool Using Markov Models

Preliminary analysis revealed strong biases in the round zero (R0) pool. Not only do base frequencies in R0 differ (19% C, 24% G, 26% A, 32%T on the reference strand), we also observed strong correlations between neighboring nucleotide positions. To account for these biases, we trained Markov Models of various order on the set of R0 reads, and tested their predictive accuracy using cross-validation on K-mer counts. We chose K = 8 because this is the longest length for which each K-mer occurs at least 100 times in R0. We found that a fifth-order Markov Model performed best ( $R^2 = 0.996$ ) when predicting 8-mer counts ([Figure S1](#)). This model was used to estimate the R0 frequencies of K-mers in all subsequent analyses.

### Determining the Effective Length of the DNA Binding Site

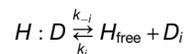
To determine the effective length of the DNA binding site, we calculated the Kullback-Leibler divergence  $D_{KL}$  for different K-mer lengths as a measure of the information gain in the pool after two rounds of affinity-based selection, relative to the fifth-order Markov Model of R0:

$$D_{KL} = \sum_{w \in S_{100}} \left( P_2(w) \log \frac{P_2(w)}{P_0(w)} \right) + \left[ 1 - \sum_{w \in S_{100}} P_2(w) \right] \log \left( \frac{1 - \sum_{w \in S_{100}} P_2(w)}{1 - \sum_{w \in S_{100}} P_0(w)} \right)$$

Here,  $P_2(w)$  represents the normalized frequency of K-mer  $w$  in R2, and  $P_0(w)$  the expected frequency of  $w$  in R0 as computed using the Markov Model. Sums are over the set  $S_{100}$  of all K-mers seen at least 100 times in R2 (corresponding to a sample error in the fold-enrichment of at most 10%). All remaining K-mers were treated as a low-affinity single category when computing  $D_{KL}$ .

### Inferring Relative Affinities for all K-mers

Our analysis of the SELEX read counts is based on a thermodynamic description of the affinity-based selection process. If  $D_i$  denotes the  $i$ -th species of DNA molecules in the pool,  $H$  the Hox-Exd complex, and  $H:D_i$  the DNA-bound complex, we have the following set of coupled equilibria:



Let the fraction of  $D_i$  in the pool be  $F_i = [D_i] / [D_{\text{tot}}]$ , and let  $K_d(D_i) = k_{-i} / k_i$  denote the dissociation constant for the  $i$ -th equilibrium. It can be shown that the post-selection frequencies  $F'$  are related to the preselection frequencies  $F$  by the following equation (Djordjevic et al., 2003; Levine and Nilsen-Hamilton, 2007):

$$\frac{F'_i}{F'_j} = \frac{(K_d(D_i) + [H_{\text{free}}]) F_i}{(K_d(D_j) + [H_{\text{free}}]) F_j}$$

Iterating this equation over multiple rounds, assuming that most of the protein complex  $H$  is bound to DNA (i.e.,  $[H_{\text{free}}] < < K_{\text{opt}}$ ) yields the following expression for the relative affinity  $k_a$  of DNA sequence  $D_i$  in terms of the frequencies in round  $r$  (Rr) and round zero (R0):

$$k_a(D_i) = \frac{K_{d,\text{opt}}}{K_d(D_i)} \approx \left( \frac{F'_i / F'_{\text{opt}}}{F_i / F_{\text{opt}}} \right)^{1/r}$$

Here,  $F_{\text{opt}}$  denotes the frequency of the highest-affinity sequence. In other words, all affinities are normalized to the interval between zero and unity. We were interested in inferring a table of relative affinities  $k_a(w)$  for all K-mers  $w$  of a given length  $K$ . Formally, this requires a deconvolution of the total affinity of each DNA molecule in terms of all ways in which it can be bound by  $H$  over a stretch of  $K$  base pairs, and in either direction. A fully systematic approach to this problem will be presented elsewhere (Riley et al., unpublished data). However, an approximate solution—which assumes that a single K-mer dominates the rate at which each DNA molecule is selected—is to adapt the above equation to the level of K-mers. Estimating the frequency  $F_w$  of DNA molecules containing a specific K-mer  $w$  as proportional to the read counts  $N_w$  yields:

$$k_a(w) \approx \left( \frac{N_w / F_{\text{opt}}^r}{P_0(w) / P_0(w_{\text{opt}})} \right)^{1/r}$$

As above,  $P_0(w)$  denotes the expected frequency of  $w$  in R0 as computed using a (fifth-order) Markov Model. The standard error of  $k_a(w)$  is dominated by the Poisson standard error of the count in the later round. The error in the Markov-Model estimate is expected to be much smaller; however, to be conservative, we assume it to be of the same order:

$$SE(k_a(w)) = k_a(w) \sqrt{\frac{2}{N_w}}$$

### LOESS-based Integration of Multiple Rounds of SELEX

For each Exd-Hox protein complex we compared the fold-enrichment from R0 to R1 with the  $n^{\text{th}}$  root of the fold-enrichment from R0 to Rn for all 12-mers. We observed a consistent deviation from a straight line, which is presumably due to a combination of binding saturation and PCR bias (Figure S1). This effect is less severe in the earlier rounds, and therefore we concluded that R1/R0 is the most

accurate predictor of relative affinity. However, since counts are lower in R1 than in subsequent rounds, the value of R1/R0 is also less precise. To leverage the more accurate counts from R1 and the more precise counts from a later round, we integrate information from multiple rounds. We assume that the fold-enrichments in the later rounds depend monotonically on the affinity. Therefore the affinities computed as  $(R_n/R_0)^{1/n}$  will have been corrected for any non-linear bias after LOESS regression on R1/R0 (Figure S1). This results in an estimate of relative affinity that is both accurate and precise. To optimize the parameters for the LOESS regression, we compared the corrected relative affinities to EMSA gel shifts results (Figure 2). An optimal fit was obtained using a 2<sup>nd</sup> order polynomial with a smoothing span of 0.2. Additionally, we found an improvement by using the R1/R0 relative affinities themselves as weights in the polynomial fit to compensate for the unequal distribution of data points. The final affinity tables presented in this paper were based on the integration of 12-mer R1/R0 enrichments with either R3/R0 enrichments (Exd-Pb and Exd-Scr) or R2/R0 enrichments (all other Exd-Hox heterodimers). The final monomer affinity tables were either obtained from an integration of 9-mer R1/R0 enrichments with either R2/R0 enrichments (Lab) or from R1/R0 enrichments alone (all other Hox monomers). These rounds were chosen to optimize counts (and thereby minimize the sampling error) over up to two orders of magnitude of relative affinity.

### Sequence Logos

The sequence logos are based on a positional-independence model, where we assume that the free energy contribution for each position in the binding site are independent and additive. Within this framework, the height of each nucleotide letter is made proportional to its relative affinity at each position in the binding site, and the letters are sorted in descending affinity order. The height of the entire stack at each position is then adjusted to signify the information content (in bits) of that position (Schneider and Stephens, 1990). The positional-independence models were generated by looking up the relative affinities for all single point mutations away from the highest-affinity consensus site, and the sequence logos were generated using BioJava 1.6 (Holland et al., 2008).

### Analysis of ChIP-Chip Data

We processed ChIP-chip data for Ubx and Hth (Slattery et al., 2011) using MAT (Johnson et al., 2006), calling peaks at a 5% false discovery rate. Genomic sequences bound by both Ubx and Hth over at least 100 base pairs were defined as “Ubx+Hth” peaks. Genomic DNA sequences were downloaded from flybase.org. The 12-mer affinity tables derived from the SELEX-seq data for each Exd-Hox were filtered to include only sequences of type nnnAYnnAYnnn. Using the sequence underlying each set of peaks, a total affinity statistic was computed by looking up all 12-bp sequences in a sliding window along both strands, and summing the corresponding affinities. To construct a null model, we extracted control sequences at offsets of -10kb, -5kb, -1kb, +1kb, +5kb, and +10k relative to each ChIP peak. We randomly selected one control window per ChIP peak and computed the total affinity statistic. This was repeated 1,000 times. The resulting null distribution was reasonably close to Gaussian, so we summarized it by its mean and standard deviation, and computed P-values using the cumulative normal distribution. Fold-enrichment was defined as the ratio of the total affinity of the peak sequences divided by the mean total affinity in the random samples. The standard error in the fold-enrichment was based on the same Gaussian approximation to the null distribution. For color-specific analyses, we required that the core be a specific hexamer (see Figure 2C for color definitions), and allowed four possible Exd flanks (nTG, nTA, nAG, nTT) on the Exd side of the binding site.

### DNA Shape Prediction

All-atom Monte Carlo (MC) simulations without the protein present were used to predict structural features intrinsic to the base sequence of the DNA targets. The MC simulations were initiated from ideal B-DNA structures of 20-mers that have the nTGAYNNAYnnn motif in the center of the variable 16-base pair region (excluding reads with more than one motif). The MC simulation protocol was described previously (Joshi et al., 2007). The sampling algorithm is based on collective and internal variables (Rohs et al., 2005), an analytic chain closure using associated Jacobians (Sklenar et al., 2006), explicit sodium counter ions, and an implicit solvent model described by a distance-dependent sigmoidal dielectric function (Rohs et al., 1999). The Metropolis-Boltzmann criterion was applied based on energy calculations within the framework of the AMBER94 force field (Cornell et al., 1995). Resulting MC trajectories were analyzed with CURVES (Lavery and Sklenar, 1989) in the TGAYNNAY direction, thereby providing average structural parameters. Independent MC simulations were performed in cases where force field artifacts led to deformations thus restricting the conformational search to B-DNA.

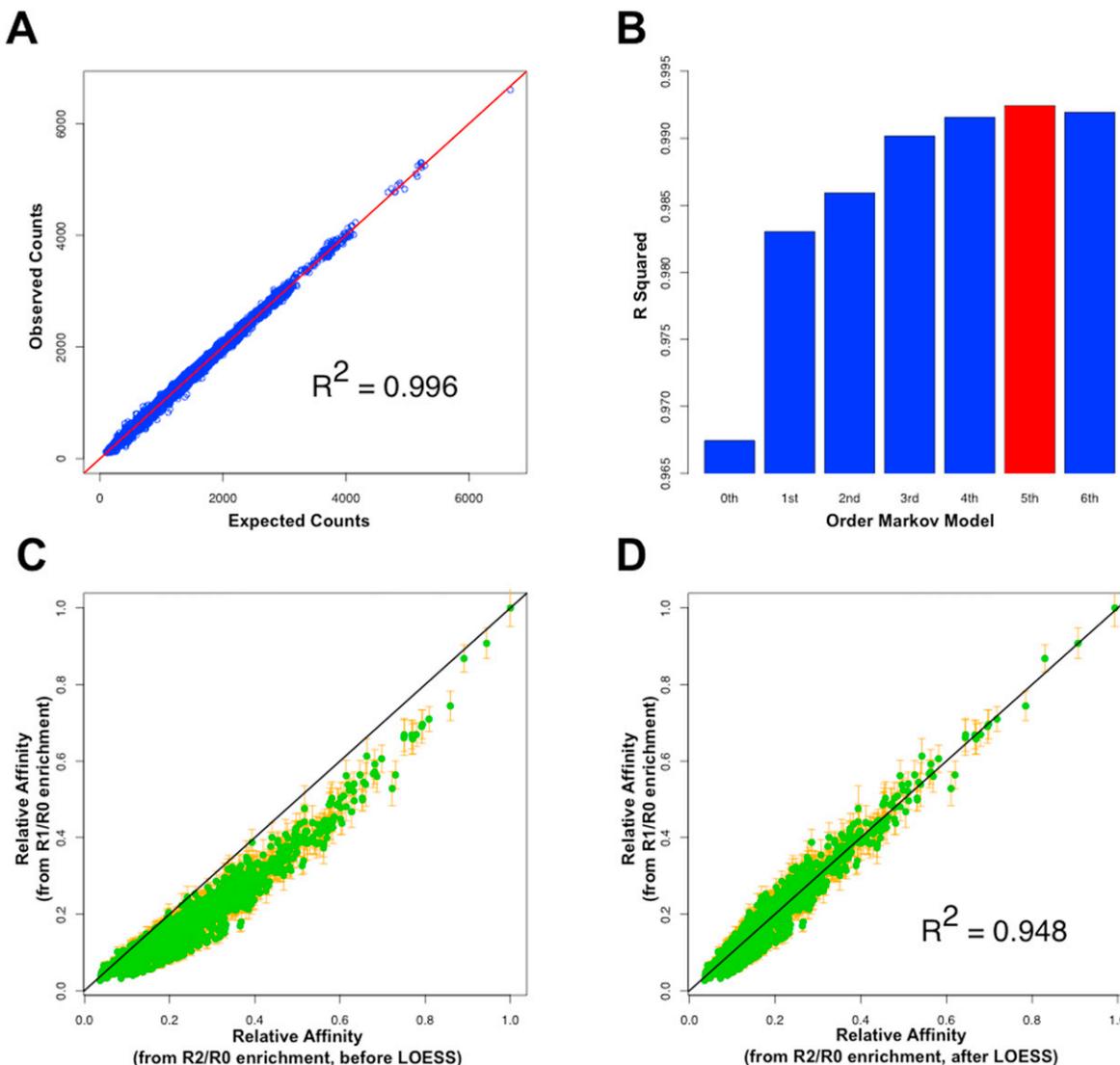
For the high-throughput analysis, a total of 1,658 trajectories from independent MC simulations were used to build a database for DNA shape predictions. These MC trajectories were analyzed in terms of the conformation of all associated tetra- and penta-nucleotides. The data derived from tetramer and pentamer conformations were combined in a hybrid model, which uses only pentamer data if the pentamer occurrence > 3, a combination of penta- and tetramer data if the pentamer occurrence ≤ 3, and only tetramer data if the pentamer occurrence is 0. The hybrid model is necessary because only 467 of the 512 unique pentamers (91%) occur in our current dataset compared to the almost complete coverage of 135 of the 136 unique tetramers. Each tetra-nucleotide occurs on average 178 times and each penta-nucleotide on average 50 times in the MC data used for the predictions. A more complete description of this method will be published elsewhere (Zhou T, Dror I, and Rohs R, unpublished).

Applying this method to the SELEX-seq binding sites, the average minor groove width at the two central nucleotides of tetramers and the central nucleotide of pentamers were used to infer the shape of all sequences that had a relative affinity above 0.1. All reads were aligned based on the TGAYNNAY motif (excluding reads with more than one motif) and the average minor groove width in each

position was calculated. We used box plots to compare differences in minor groove width at the most distinct positions A<sub>8</sub> and Y<sub>9</sub> and calculated Mann-Whitney U p-values to show the significance of differences in shape between the two groups, class 1+2 and class 3 Exd-Hox sites. To further evaluate the similarities between the different Exd-Hox sites, we compared the average width in all positions of the 12-mer nTGAYNNAYnnn using Pearson correlation. The width values at the six positions of the AYNAY core motif were used to calculate a Euclidean distance tree that relates the shapes selected by all Exd-Hox dimers. This dendrogram was generated with the UPGMA method as implemented in the MEGA program (Tamura et al., 2011).

#### SUPPLEMENTAL REFERENCES

- 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.
- Chan, S.K., Popperl, H., Krumlauf, R., and Mann, R.S. (1996). An extradenticle-induced conformational change in a HOX protein overcomes an inhibitory function of the conserved hexapeptide motif. *EMBO J.* *15*, 2476–2487.
- Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.J., Ferguson, D.M., Spellmeyer, D., Fox, T., Caldwell, J., and Kollman, P.A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* *117*, 5179–5197.
- Djordjevic, M., Sengupta, A.M., and Shraiman, B.I. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Res.* *13*, 2381–2390.
- Gebelein, B., Culi, J., Ryoo, H.D., Zhang, W., and Mann, R.S. (2002). Specificity of Distalless repression and limb primordia development by abdominal Hox proteins. *Dev. Cell* *3*, 487–498.
- Holland, R.C., Down, T.A., Pocock, M., Prlic, A., Huen, D., James, K., Foisy, S., Drager, A., Yates, A., Heuer, M., et al. (2008). BioJava: an open-source framework for bioinformatics. *Bioinformatics* *24*, 2096–2097.
- Jaffe, L., Ryoo, H.D., and Mann, R.S. (1997). A role for phosphorylation by casein kinase II in modulating Antennapedia activity in *Drosophila*. *Genes Dev.* *11*, 1327–1340.
- Johnson, W.E., Li, W., Meyer, C.A., Gottardo, R., Carroll, J.S., Brown, M., and Liu, X.S. (2006). Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl. Acad. Sci. USA* *103*, 12457–12462.
- 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.
- 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.
- Lavery, R., and Sklenar, H. (1989). Defining the structure of irregular nucleic acids: conventions and principles. *J. Biomol. Struct. Dyn.* *6*, 655–667.
- Lefrancois, P., Euskirchen, G.M., Auerbach, R.K., Rozowsky, J., Gibson, T., Yellman, C.M., Gerstein, M., and Snyder, M. (2009). Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genomics* *10*, 37.
- Levine, H.A., and Nilsen-Hamilton, M. (2007). A mathematical analysis of SELEX. *Comput. Biol. Chem.* *31*, 11–35.
- 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.
- 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.
- 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.
- Rohs, R., Etchebest, C., and Lavery, R. (1999). Unraveling proteins: a molecular mechanics study. *Biophys. J.* *76*, 2760–2768.
- 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.
- Ryoo, H.D., and Mann, R.S. (1999). The control of trunk Hox specificity and activity by Extradenticle. *Genes Dev.* *13*, 1704–1716.
- Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* *18*, 6097–6100.
- Sklenar, H., Wustner, D., and Rohs, R. (2006). Using internal and collective variables in Monte Carlo simulations of nucleic acid structures: chain breakage/closure algorithm and associated Jacobians. *J. Comput. Chem.* *27*, 309–315.
- 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.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* Published online: 10.1093/molbev/msr121.



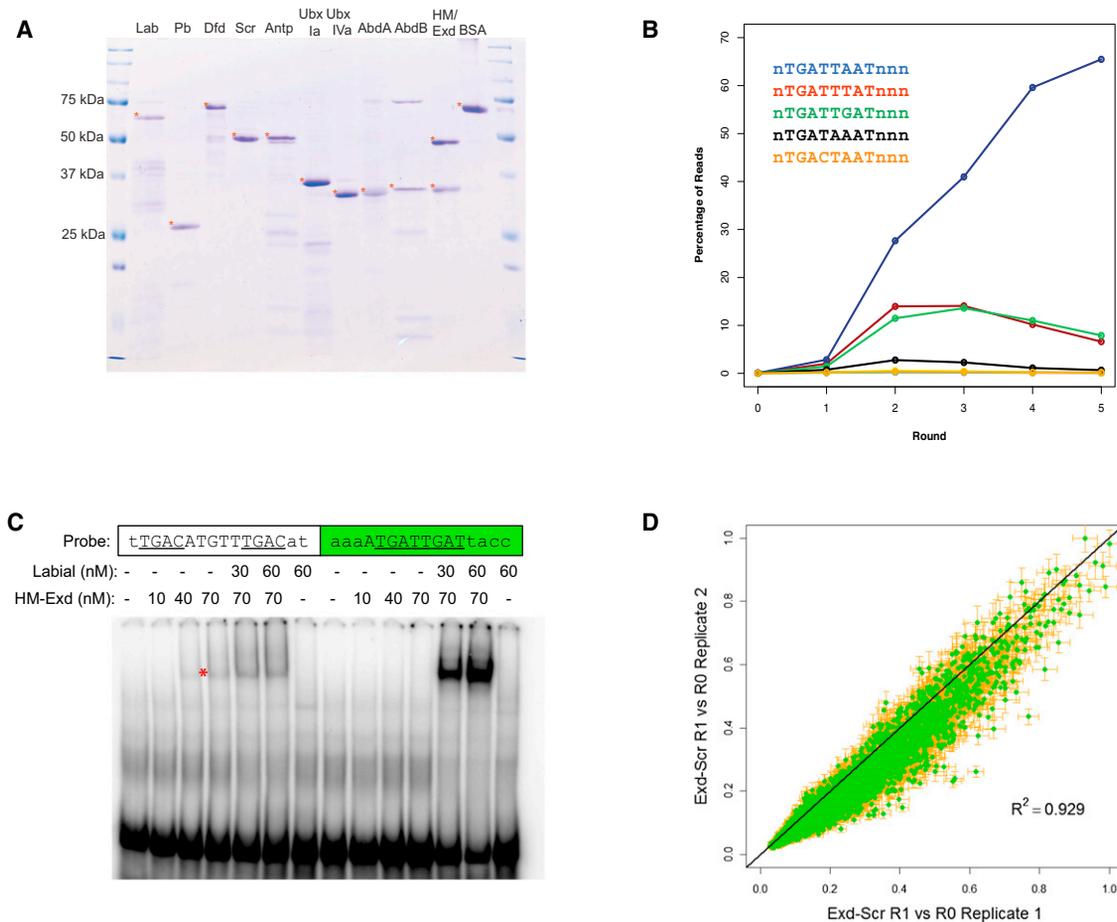
**Figure S1. Detailed Methodology for Inferring Relative Affinities for all K-mers, Related to Figure 1**

(A) Although the DNA library was synthesized by “hand-mixing” (Integrated DNA Technologies), sequencing of R0 revealed significant biases in the  $N_{16}$  region, likely due to biases in the synthesis of the initial library and the Klenow reaction required to make the DNA double stranded. We account for the biases present in the initial pool (R0) by parameterizing the relative frequency of each 16-mer using a standard Markov model. To validate this approach, we split the reads into two subsets based on the multiplex tags shown in Table S1. We trained a fifth-order Markov model on one subset, and used it to predict the frequency of all possible 8-mers in the other subset. Shown is a direct comparison between predicted and observed 8-mer counts in the test set. While 8-mer frequencies vary over almost three orders of magnitude in R0 (poly-T being the most, and poly-C motifs the least abundant), the Markov model does an excellent job capturing this variation (adjusted  $R^2 = 0.996$ ).

(B) We compared the values of  $R^2$  for Markov models of order zero through six, and found that a fifth-order model has the best cross-validation performance. At lower order, the biases in R0 are not sufficiently captured; at higher order, the predictions degrade due to over-fitting.

(C) To compute relative affinities for all 12-mers, we integrate information from multiple rounds of selection using LOESS regression. Shown is a direct comparison between the fold-enrichment from R0 to R1 and the square root of the fold-enrichment from R0 to R2 for all 12-mers during in vitro selection for binding by the Exd-Lab heterodimer. The deviation from the straight line is presumably due to a combination of binding saturation and PCR bias. These effects are expected to be less severe in the earlier round, and therefore R1/R0 is a more accurate predictor of relative affinity. However, since counts are lower in R1 than in R2, the value of R1/R0 is also less precise. 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) After LOESS regression on R1/R0, the affinities computed as  $(R2/R0)^{1/2}$  have been corrected for saturation and PCR bias (adjusted  $R^2 = 0.948$ ). This results in an estimate of relative affinity that is both accurate and precise. The error bars denote the standard error in the estimate of the relative affinity as calculated based on Poisson statistics (see [Extended Experimental Procedures](#)).



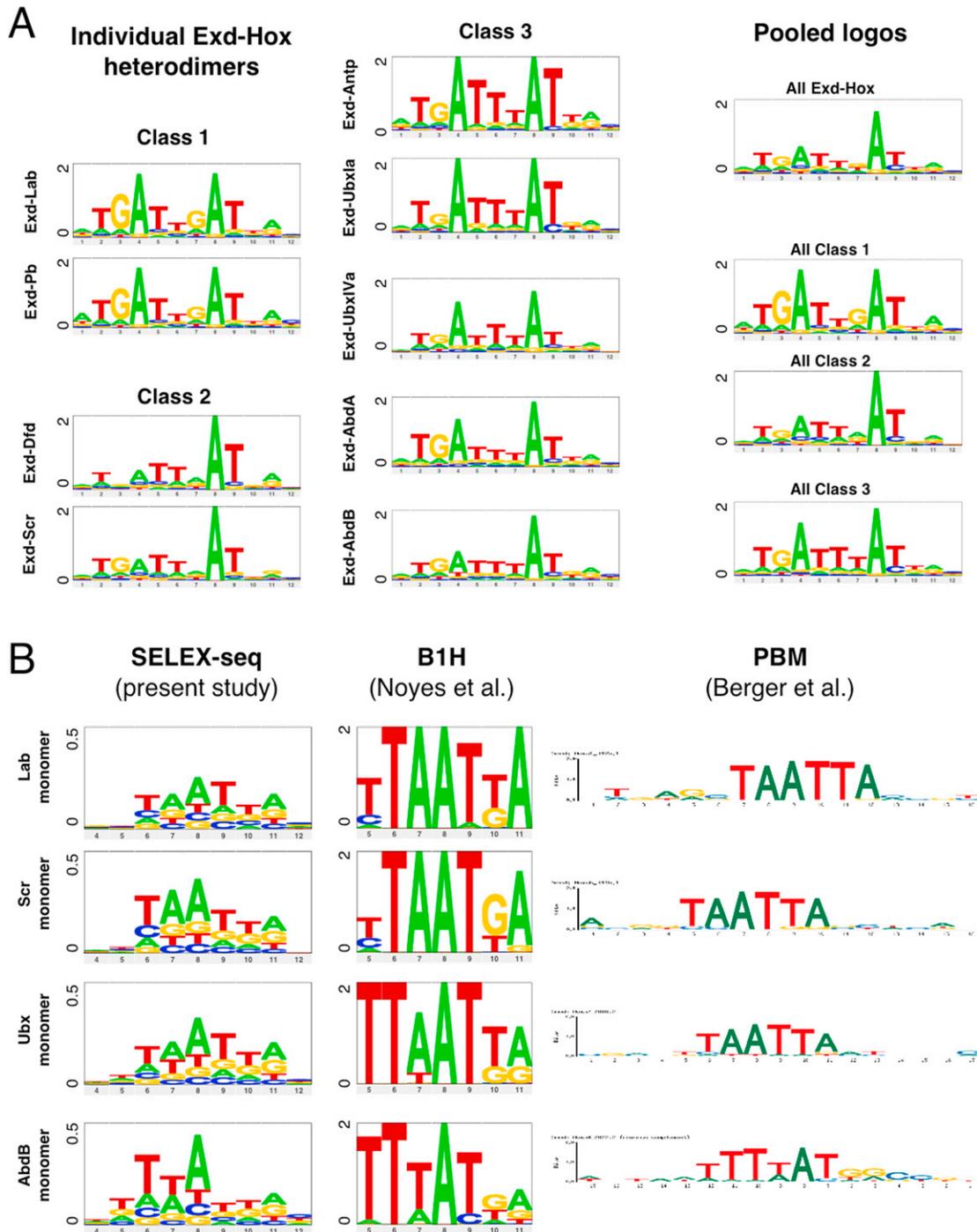
**Figure S2. EMSA and SELEX Quality Control, Related to Figure 2**

(A) The indicated affinity-purified recombinant proteins and BSA were resolved by SDS-PAGE and visualized by staining with Coomassie Blue. In all cases the recombinant protein is indicated with a red asterisk; for HM-Exd, both proteins can be visualized at approximately equal stoichiometries.

(B) Evolution of DNA pool composition over multiple rounds of selection for binding by Exd-Scr. For each round of selection, the plot shows the percentage of reads that contain an Exd-Hox motif of a particular color. As expected for a class 2 Hox protein, DNA molecules containing the blue motif are selected at the highest rate. DNA molecules containing the red and green motifs are also selected, but at a lower rate due to the lower affinity with which they are bound by Exd-Scr. In later rounds, they are outcompeted by the blue motif. Our relative affinity calculations for 12-mers for Exd-Scr integrate information from R0, R1, and R2.

(C) Analyzing Exd-Hox heterodimer versus Exd-Exd or Hox-Hox homodimer binding. Hox-independent Exd-Exd binding runs at the same mobility as an Exd-Hox binding event in an EMSA. Significant Hox-independent HM-Exd binding is observed with the sequence on the left, which matches the “Exd-Exd” sequences identified in SELEX and described in panel A (TGAY(N<sub>5</sub>)TGAY, in this case). This Exd-Exd complex, indicated with a red asterisk, travels at nearly the same mobility as the cooperative Exd-Lab complex.

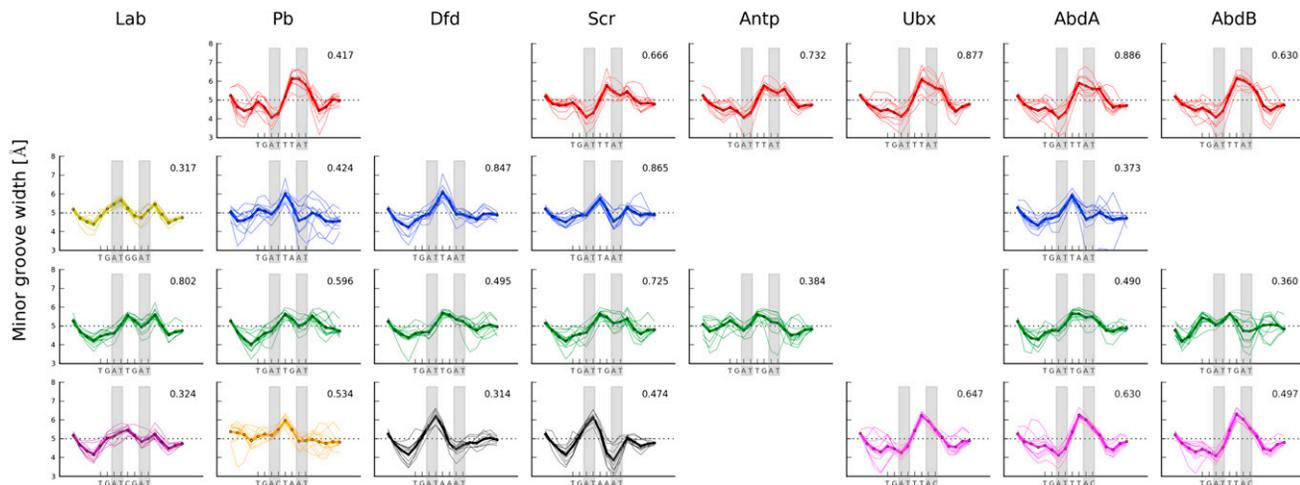
(D) Reproducibility of heterodimeric Exd-Scr relative affinities from two independent replicates after one round of SELEX. Each plot is a direct comparison between 12-mer affinities estimated as relative R0⇒R1 enrichments. Both the protein preps and SELEX for these two replicates were independent. The error bars denote the standard error in the estimate of the relative affinity as calculated based on Poisson statistics (see [Extended Experimental Procedures](#)).



**Figure S3. Sequence Logos Derived from SELEX-Seq Data, Related to Figures 3 and 4**

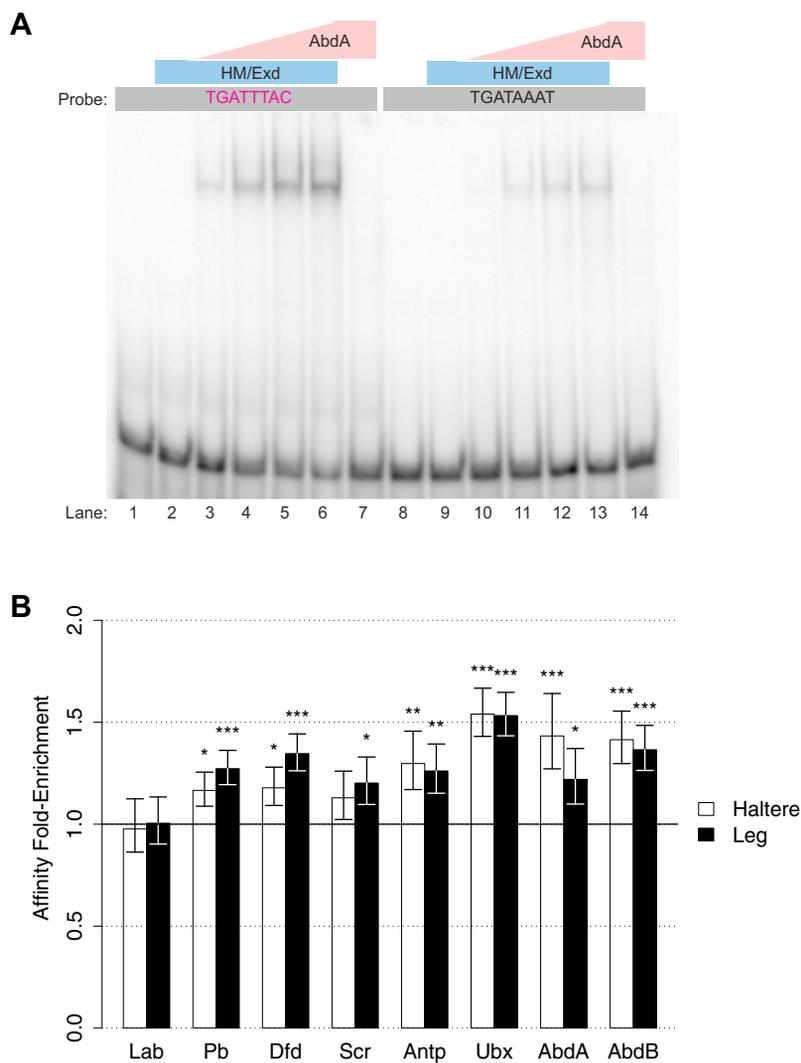
(A) Logos for each of the nine Exd-Hox complexes, organized by class, as derived from the 12-mer tables of relative affinities obtained by SELEX-seq. Also shown is a “Hox-blind” consensus Exd-Hox logo, in which positions 6 and 7 are the most variable as expected, as well as class-specific consensus logos.

(B) Comparison of sequence logos for Hox monomers obtained using different technologies. Logos derived from 9-mer tables of relative affinities obtained by SELEX-seq (this study) are shown alongside those derived using bacterial 1-hybrid technology (B1H; Noyes et al., 2008) and protein binding microarrays (PBM; Berger et al., 2008), respectively. The results are qualitatively consistent across technologies, and show that differences in specificity elicited by heterodimerization with Exd are absent for Hox monomers. Together, the four Hox proteins shown here span the three Specificity Classes for Exd-Hox dimers (Class 1: Lab; Class 2: Scr; Class 3: Ubx and AbdB). Note that the overall information content in the SELEX-seq logos is lower, which might be a consequence of the larger number of sequences that were used to generate them, and our ability to resolve low relative binding affinities. In addition, unlike the earlier studies, the proteins in the present study have more than just the DNA binding domain and are much closer to full-length.



**Figure S4. Predicted Minor Groove Widths of Exd-Hox Binding Sites, Related to Figure 6**

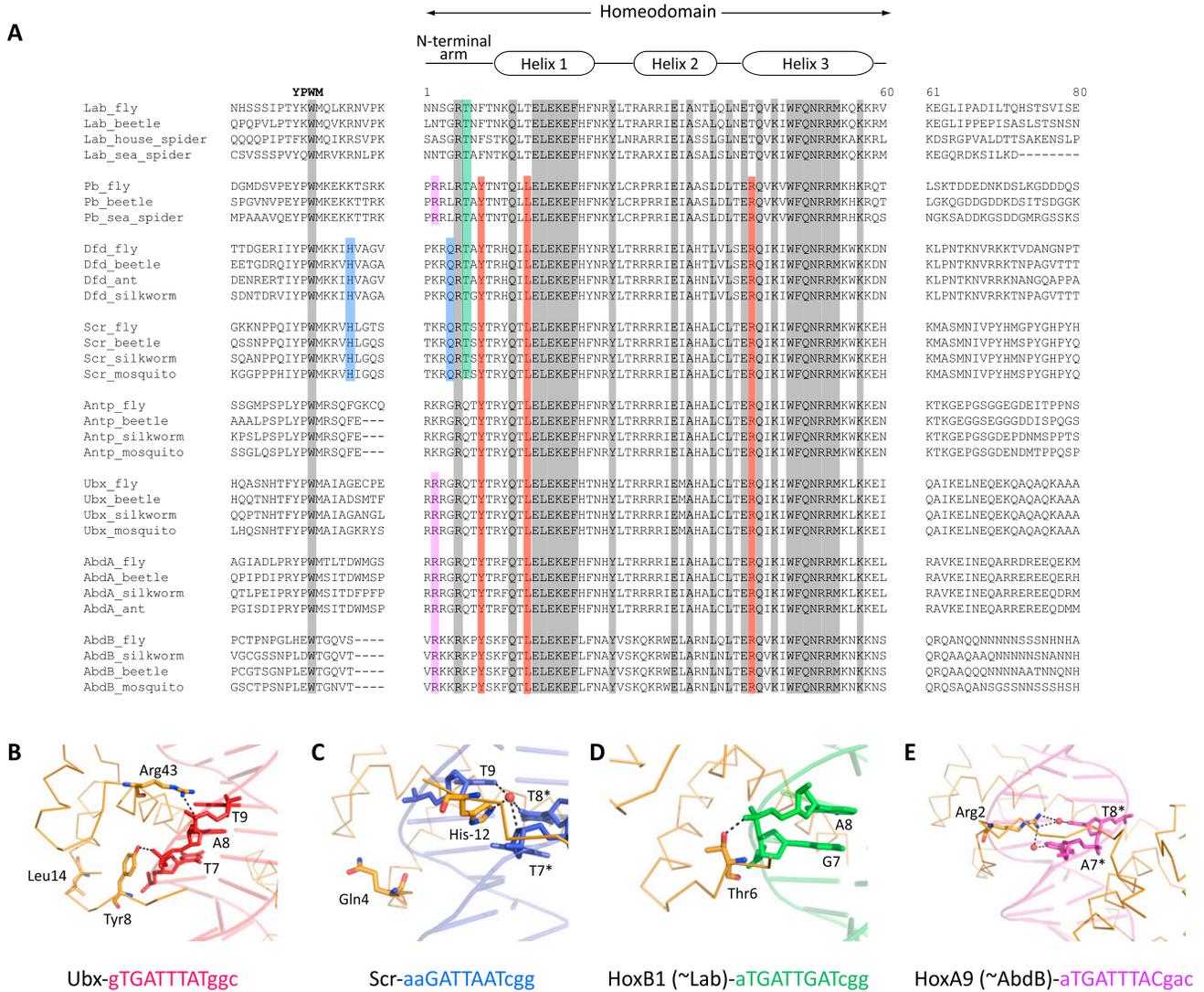
Shown are Monte Carlo predictions of minor groove width for the ten highest affinity binding sites for each Exd-Hox complex (thin lines) complemented by the average prediction for each motif (thick lines). Up to four core motifs for a given complex were included in this analysis if their average relative affinity for the top ten binders was above 0.3. Hox protein identity is denoted on top of each column (Ubx represents isoform Ia). The core sequence of the DNA binding site in each graph is listed below the x axis, and the relative average binding affinity is indicated in the top right of each graph. These plots illustrate that most sequences have minima in the A<sub>4</sub>T<sub>5</sub> region, which extend in the Exd direction, probably due to the presence of short A-tracts in many of the sequences. This region likely accommodates the conserved Arg5 residues of both Exd and Hox. The largest variation between these binding sites is apparent in the A<sub>8</sub>T<sub>9</sub> region. This difference originates from replacing a purine at position 7 (an A in about 47% and a G in about 48% of class 1 and 2 sites above a relative affinity of 0.1) with a T (in 79% of class 3 sites above 0.1 relative binding affinity), which shifts the location of a TpR step in 3' direction. Notably, the replacement of an A with a G at position 7 forms a CpA step on the opposite strand, which has similar properties to the TpA step, thus accounting for the presence of either a TpA or TpG step in class 2 proteins. While a more detailed understanding of the role of positions 6 and 7 will benefit from additional crystal structures, the shape analysis presented here nevertheless highlights the general importance of DNA shape for specific DNA binding by Hox proteins.



**Figure S5. In Vitro and In Vivo Confirmation of Exd-Hox DNA Binding Preferences, Related to Figure 7**

(A) Exd-AbdA preferentially binds magenta motifs over black motifs. Exd-AbdA binding to a black motif (lanes 8-14) is significantly weaker than Exd-AbdA binding to a magenta motif (lanes 1-7). The full magenta sequence is 5'-CAAACCCAGTTCAGAGCGAATGATTTACGACCGGTCAAGGTCGTTTCC and the full black sequence is 5'-CAAACCCAGTTCAGAGCGAATGATAAATGACCGGTCAAGGTCGTTTCC.

(B) Comparing in vivo and in vitro binding by Exd-Ubx. We tested whether the predicted affinity for each Exd-Hox complex (based on the SELEX-seq data) is associated with in vivo Exd-Ubx binding in the T3 leg and haltere imaginal discs previously identified using ChIP-chip (Slattery et al., 2011). We computed the total affinity per kb across the set of all genomic windows (median size 350 bp) that were occupied both by Ubx and Hth (an obligatory binding partner of Exd). Shown is the fold-enrichment of this affinity density over a set of control regions taken from the flanking regions of the binding sites. As expected, the highest enrichment occurs when affinity is predicted using the sequence-to-affinity model for Exd-Ubx, rather than for one of the other Exd-Hox heterodimers. The symbols above each bar denote the statistical significance level (\*\*\* $p \leq 0.001$ , \*\* $p \leq 0.01$ , \* $p \leq 0.05$ ). Error bars correspond to standard errors, computed based on a thousand samples from the control distribution.



**Figure S6. Hox Protein Sequences Correlate with DNA Binding Specificity, Related to Figure 6**

(A) Alignments of sequences surrounding the homeodomains and YPWM motifs of Hox proteins from arthropods. (B-E) Existing PBC-Hox X-ray crystal structures (PDB IDs: (B) 1B8I (Passner et al., 1999); (C) 2R5Z (Joshi et al., 2007); (D) 1B72 (Piper et al., 1999); (E) 1PUF (LaRonde-LeBlanc and Wolberger, 2003)). Class I is represented in (D) by HoxB1 (orthologous to Lab); Class II is represented in (C) by Scr; and Class III is represented in (B) by Ubx and (E) by HoxA9 (orthologous to AbdB). The name of the Hox protein and its 12-mer DNA binding sites are listed below each panel. Partially conserved Hox residues that correlate with DNA binding specificity as highlighted by the colors in (A) and are labeled and shown in sticks and highlighted in the structures in B-E. Nucleotides that form direct or water-mediated hydrogen bonds with these residues are labeled and represented in sticks as well. Numbering of the labeled nucleotides is based on the 12 base pair binding site. Nucleotides on the complementary strand are labeled with stars. Hydrogen bonds are represented as dashed lines and water molecules are shown in red spheres. Methods: Sequences for Hox proteins were obtained from NCBI's Protein database (<http://www.ncbi.nlm.nih.gov/protein/advanced>). The species used were restricted to arthropods and sequences that contain more than 80 amino acids. Except for AbdB, all the Hox sequences must contain a WM motif. In cases where more than one sequence from the same species was present, the longest sequence was used. Protein sequences that contain "predicted," "putative," or "hypothetical" in the headlines of their fasta files were discarded. Homeodomains were identified through the conserved Q50-N51-R52-R53 motif and the YPWM motif was determined by the conserved tryptophan residue across all eight Hox proteins. Partially conserved residues that correlate with DNA-binding specificity were identified by comparing the multiple sequence alignment with Figure 3A. For each of the four most favored hexamer DNA binding sites (red, blue, green, and magenta motif in Figure 3A), a subset of Hox proteins was built to include those having relative binding affinity higher than 0.5. Then, for each aligned position, residues that are 100% conserved across Hox proteins in the subset, but have different residue types in the other Hox proteins, were defined as the partially conserved residues that correlate with DNA-binding specificity. Exd-Pb's hexamer preference was not considered during the identification of partially conserved residues. Hydrogen bonds between partially conserved residues and DNA were from contact maps in the papers that describe the crystal structures (Joshi et al., 2007; LaRonde-LeBlanc and Wolberger, 2003; Passner et al., 1999; Piper et al., 1999).