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), UbxIa (Ryoo and Mann, 1999), UbxIVa (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 μl. 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 μl 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 μl) was then amplified by PCR with the primers "SELEX SR 0" and "SELEX SR 1." For PCR, the 10 μl of eluted DNA was split equally among five, 50 μl reactions (0.4 μM 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 μl reactions with 0.8 μM 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:

$$H : D \underset{k_i}{\overset{k_{-i}}{\rightleftarrows}} H_{\text{free}} + D_i$$

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):

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

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):

$$k_a(D_i) = \frac{K_{d,\text{opt}}}{K_d(D_i)} \approx \left(\frac{F_i^r / F_{\text{opt}}^r}{F_i^0 / F_{\text{opt}}^0} \right)^{1/r}$$

Here, F_{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:

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

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:

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

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^{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 $(Rn/R0)^{1/n}$ 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^{nd} 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_8 and Y_9 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 AYNNAY 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.

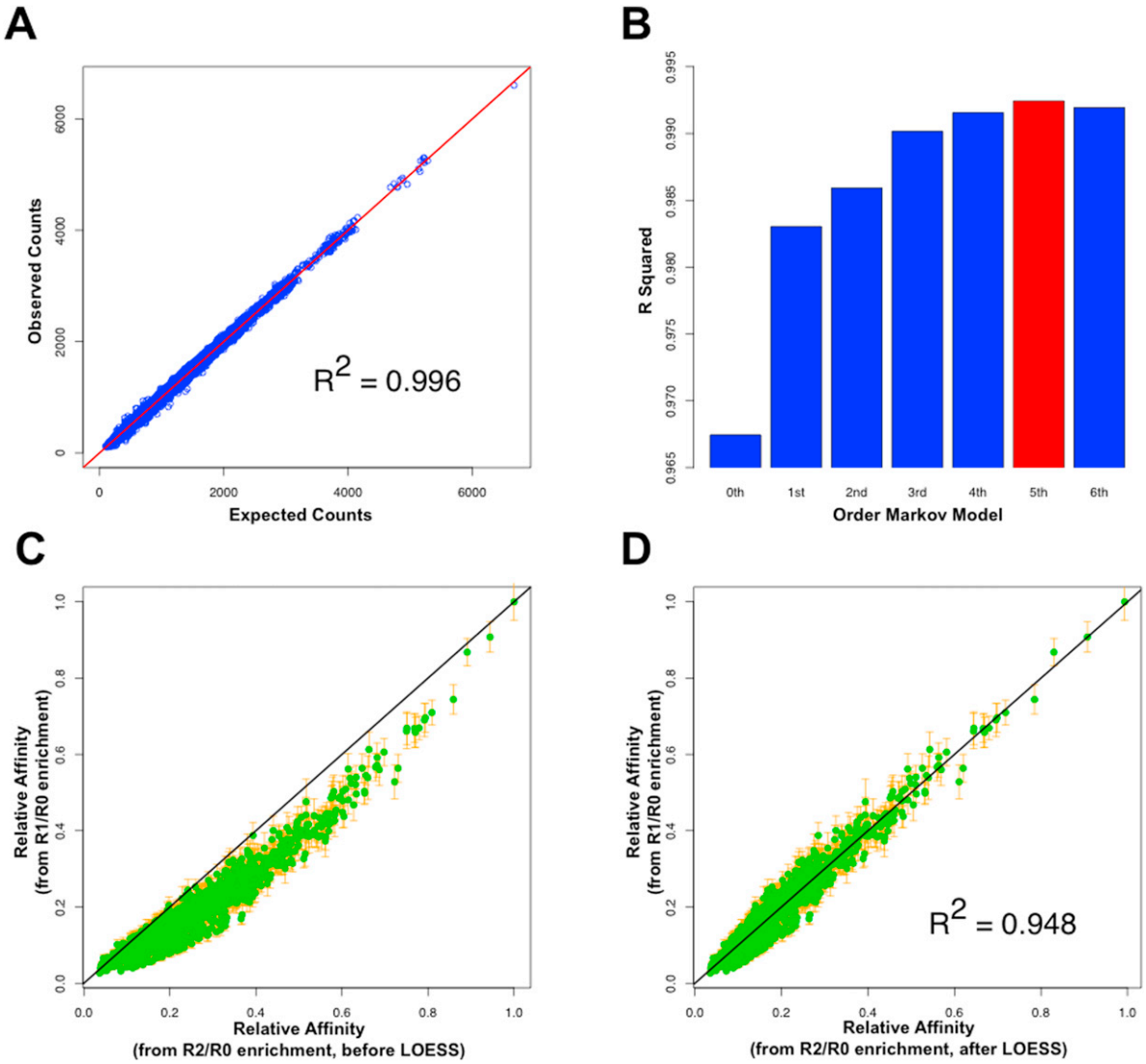

Figure S1. Detailed Methodology for Inferring Relative Affinities for all K-mers, Related to Figure 1

(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 in (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).

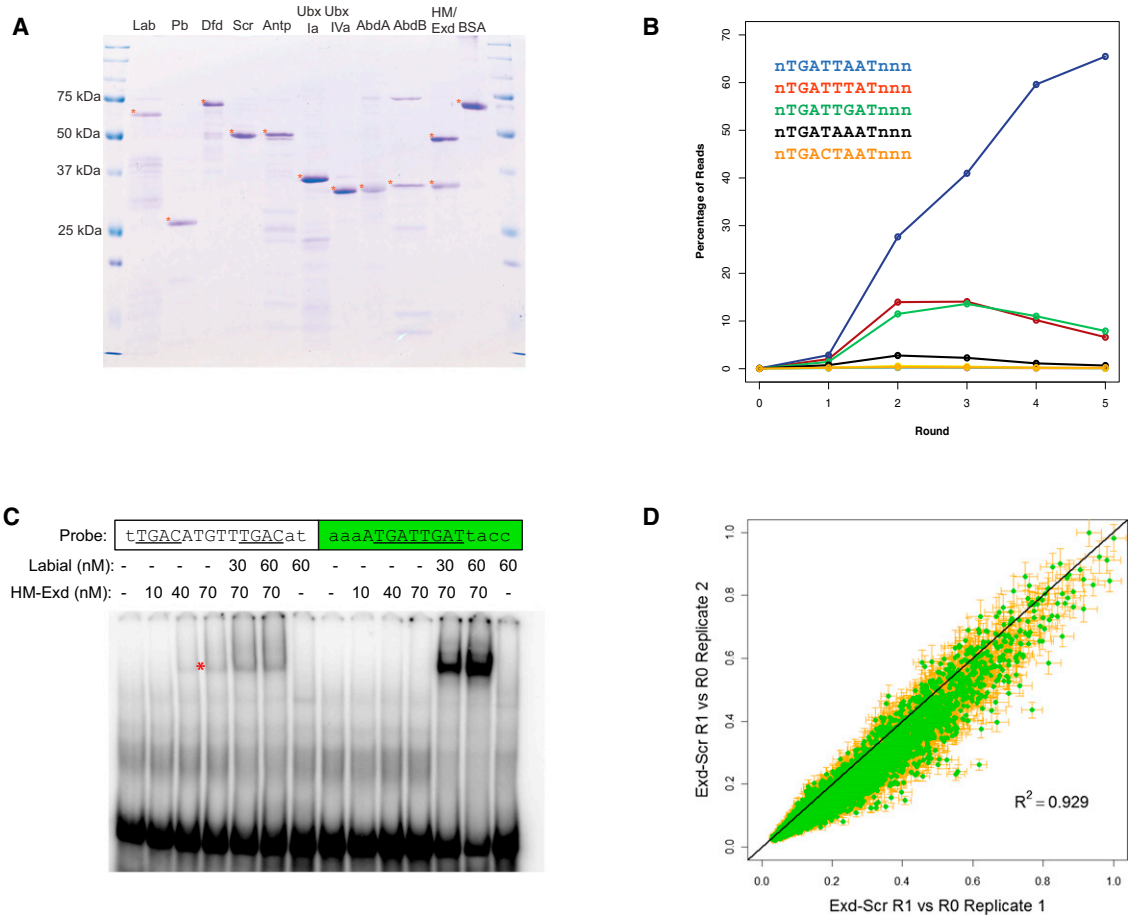

Figure S2. EMSA and SELEX Quality Control, Related to Figure 2

(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_5}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).

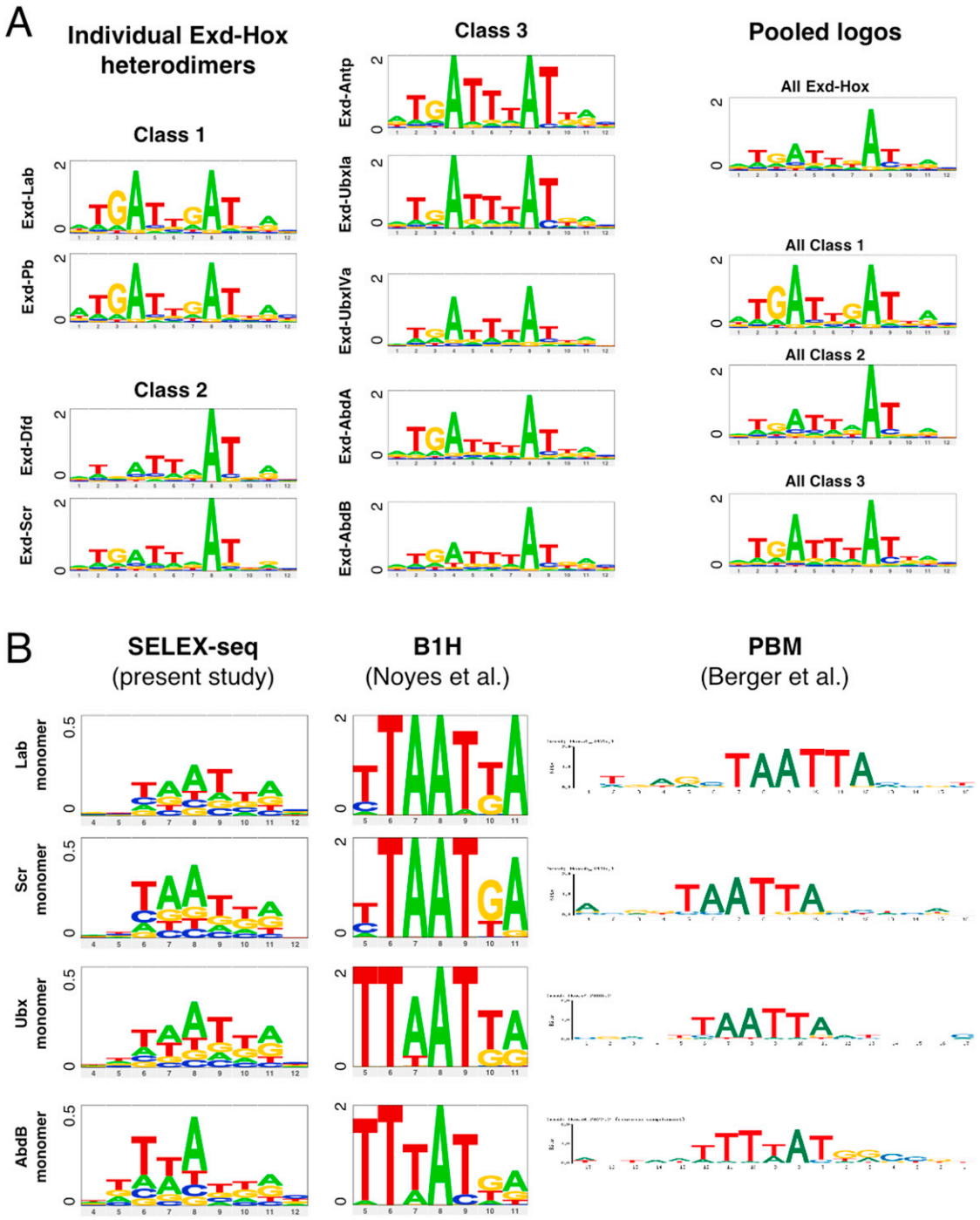

Figure S3. Sequence Logos Derived from SELEX-Seq Data, Related to Figures 3 and 4

(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.

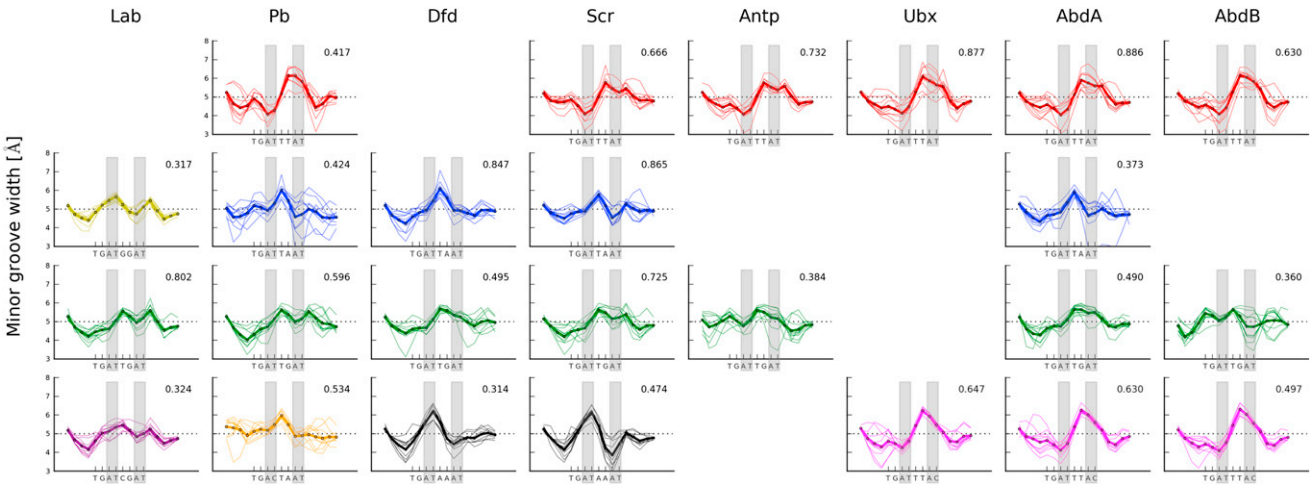

Figure S4. Predicted Minor Groove Widths of Exd-Hox Binding Sites, Related to Figure 6

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_4T_5 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_8T_9 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.

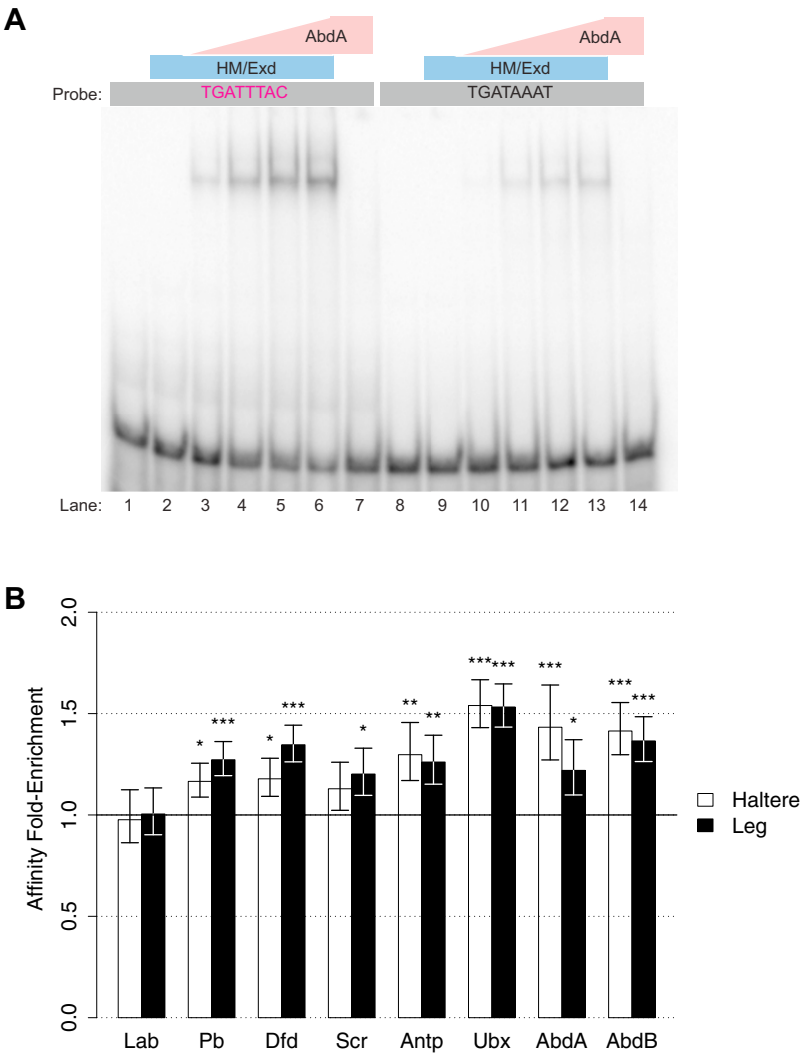

Figure S5. In Vitro and In Vivo Confirmation of Exd-Hox DNA Binding Preferences, Related to Figure 7

(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.

Slattery et al.

SUPPLEMENTARY TABLES

Supplementary Table S1: Oligonucleotides used in this study, related to Figure 1.

Supplementary Table S2: From counts to relative affinities, related to Figure 1. This table illustrates how we estimate the relative affinity of each 12-mer for Exd-Scr. First, the fifth-order Markov model derived from the R0 data is used to estimate the expected count of each 12-mer motif in the initial pool. Next, two independent estimates of the relative affinity are computed: (i) as the ratio between the R1 count and the R0 expected count, normalized (divided) by that of the highest-affinity motif; (ii) as the square root of the ratio of the R2 count and the R0 expected count, again normalized by that of the highest-affinity motif. Finally, the R2/R0 affinities are reconciled with the R1/R0 affinities using a monotonous nonlinear transformation based on LOESS regression (see Supplementary Figure S1C).

Supplementary Table S3: Top 50 dodecamers containing AYNNAY for all Exd-Hox heterodimers, related to Figure 2. The preference of anterior (Lab, Pb), central (Dfd, Scr), and posterior (Antp, Ubx, AbdA, AbdB) Hox proteins for DNA molecules containing green (ATTGAT), blue (ATTAAT), and red (ATTTAT) motifs, respectively, is readily apparent from this table. The high frequency of T-rich containing sequences in the Exd-Pb selections is likely due to the inferior nature of the Pb protein preparation, which allows PCR-based biases to compete with bona fide binding sites.

Supplementary Table S4: Composition of late-round DNA pools in terms of different core motif colors, related to Figure 3. Each group of three sequences includes tolerated variations in the Exd flank for each core. While the various Exd-Hox heterodimers select motifs of different color, most of the DNA molecules in the later rounds (R3 or R4, as indicated) contain at least one of the top 10 core motifs we identified. The gel mobility of Hox-Hox and Exd-Exd homodimers is similar to that of Exd-Hox heterodimers, and therefore these alternative complexes can also contribute to the enrichment for specific DNA sequences. Indeed, for some pools a significant fraction of the reads can be accounted for in terms of homodimer binding, based on consensus motifs that we derived for Exd-Exd and Hox-Hox complexes. This is particularly true for Pb, which selected oligonucleotides containing two Hox monomer sites, TAATTA, at high frequency. We also confirmed that sequences fitting the consensus TGAY{N_{3-}

$_{5}$}TGAY are Exd-Exd dimer binding sites by carrying out SELEX-seq experiments with HM-Exd in the absence of any Hox protein (last column); sequences fitting this consensus were the most enriched in these experiments, and more highly enriched than any of the Exd-Hox core motifs. Therefore, we can conclude from the relatively low representation of the Exd-Exd consensus in the Exd-Hox experiments that the contribution from Exd-Exd binding was limited. Note that the percentages are low in this HM-Exd-alone SELEX experiment because the data are only from a single round of selection (R1) and HM-Exd binds DNA much more poorly compared to HM-Exd-Hox complexes.

.

Supplementary Table S1, Slattery et al

Oligo Name	**Sequence**
SELEX 16mer Multiplex 1 *	5′ GTTCAGAGTTCTACAGTCCGACGATCTGG[N_{16}]CCA**gc**T**g**TCGTATGCCGTCTTCTGCTTG
SELEX 16mer Multiplex 2 *	5′ GTTCAGAGTTCTACAGTCCGACGATCTGG[N_{16}]CCA**cg**T**c**TCGTATGCCGTCTTCTGCTTG
SELEX SR 0	5′ GTTCAGAGTTCTACAGTCCGA
SELEX SR 1	5′ CAAGCAGAAGACGGCATACGA
SELEX SR 2	5′ AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA
Red_Kd	5′ CAAACCCAGTTCAGAGcgaATGATTTATgaccGGTCAAGGTCGTTTCC
Blue_Kd	5′ CAAACCCAGTTCAGAGcgaATGATTAATgaccGGTCAAGGTCGTTTCC
Green_Kd	5′ CAAACCCAGTTCAGAGaaaATGATTGATtaccGGTCAAGGTCGTTTCC
Yellow_Kd	5′ CAAACCCAGTTCAGAGaaaATGATGGATtaccGGTCAAGGTCGTTTCC
fkhCON tracking	5′ GCTATACTGTGCTATCCACAGTTCAGAGTCGTCAAGATTTATGGCCTGCTGGTCACTGGTCGTTTCCCTCTT
Lab-Exd tracking	5′ GCTATACTGTGCTATCCACAGTTCAGAGTCGCAATGATTGATCGCCTGCTGGTCACTGGTCGTTTCCCTCTT
AbdB-Exd tracking	5′ GCTATACTGTGCTATCCACAGTTCAGAGTCGCAATGATTTACGACCTGCTGGTCACTGGTCGTTTCCCTCTT
Pb-Exd tracking	5′ GCTATACTGTGCTATCCACAGTTCAGAGTCGAAAATGATTGATTACCGCTGGTCACTGGTCGTTTCCCTCTT
Exd-Exd tracking	5′ GCTATACTGTGCTATCCACAGTTCAGAGTCGTTGACATGTTTGACATGCTGGTCACTGGTCGTTTCCCTCTT

* Barcodes are lower case and bold. [N_{16}] represents the 16 randomized bases.

Supplementary Table S2, Slattery et al

	12-mer	expected R0 count (Markov Model)	R1 count	R2 count	Rel. Affinity (R1 vs R0, raw)	Rel. Affinity (R2 vs R0, raw)	Rel. Affinity (after LOESS)
1	ATGATTAATTGC	0.207	1026	22882	1.000	1.000	1.00
2	GCAATTAATCAT	0.140	640	14521	0.922	0.969	0.96
3	GTAATTAATCAT	0.223	1090	22387	0.983	0.951	0.94
4	ATGATTAATTAC	0.217	917	20776	0.851	0.930	0.91
5	GTCATTAATCAT	0.168	724	16089	0.867	0.929	0.91
6	ATGATTAATGAC	0.155	694	14626	0.902	0.923	0.91
7	AATGATTAATGA	0.201	911	18036	0.912	0.900	0.88
8	ATGATTAATGGC	0.145	585	12912	0.814	0.897	0.88
9	AATGATTAATGG	0.203	860	17623	0.855	0.886	0.86
10	AATGATTAATTG	0.266	1102	22814	0.834	0.880	0.86
11	GTAATCAATCAT	0.139	570	11803	0.823	0.874	0.85
12	GATGATTAATTG	0.238	960	20097	0.811	0.872	0.85
13	ATGATTAATGAG	0.190	840	15660	0.890	0.862	0.83
14	TAATTAATCATC	0.174	743	14028	0.860	0.853	0.82
15	TCATTAATCATT	0.225	929	18064	0.830	0.851	0.82
16	ATGATTGATTAC	0.202	793	16183	0.791	0.851	0.82
17	GATGATTAATTA	0.254	961	20322	0.763	0.850	0.82
18	GATGATTAATGA	0.180	721	14219	0.805	0.844	0.81
19	AATGATTAATTA	0.283	1188	22194	0.846	0.842	0.81
20	GATGATTAATGG	0.182	712	14205	0.790	0.840	0.81
21	TAATTAATCATT	0.293	1282	22817	0.882	0.839	0.81
22	ATGATTTATTGC	0.271	1083	20912	0.805	0.835	0.80
23	ATGATTAATGGG	0.209	799	16125	0.770	0.835	0.80
24	CAATTAATCATC	0.108	423	8254	0.786	0.829	0.80
25	ATGATTGATTGC	0.184	696	13974	0.762	0.828	0.79
26	CAATTAATCATT	0.182	664	13818	0.734	0.827	0.79
27	ATGATTAATTAG	0.266	1084	19882	0.820	0.821	0.79
28	ATGATTAATTGG	0.264	1021	19665	0.778	0.820	0.78
29	ATGATTAATGAT	0.254	976	18774	0.775	0.818	0.78
30	GCCATTAATCAT	0.096	310	7078	0.647	0.814	0.78
31	TCATTAATCATC	0.134	481	9779	0.723	0.812	0.78
32	GAATGATTAATG	0.197	774	14093	0.791	0.804	0.77
33	GCAATCAATCAT	0.090	332	6400	0.747	0.803	0.77
34	ATGATTAATGGT	0.255	978	18145	0.773	0.802	0.76
35	CCATTAATCATT	0.129	414	9105	0.647	0.798	0.76
36	GTCATCAATCAT	0.105	358	7394	0.685	0.796	0.76
37	CTCATTAATCAT	0.129	464	9024	0.726	0.796	0.76
38	ATGATTTATTAC	0.290	1115	20094	0.774	0.791	0.75
39	ATGATTGATGAC	0.139	453	9470	0.659	0.786	0.75
40	GCAATAAATCAT	0.113	394	7645	0.702	0.782	0.74
41	GTAATAAATCAT	0.179	672	12066	0.756	0.780	0.74
42	ATCATTAATCAT	0.178	614	11836	0.694	0.775	0.73
43	ATGATTAATTGT	0.338	1257	22403	0.750	0.774	0.73
44	ATGATTAATAGC	0.159	569	10543	0.720	0.773	0.73
45	CCCATTAATCAT	0.074	241	4921	0.652	0.773	0.73
46	CCATTAATCATC	0.077	279	5064	0.733	0.772	0.73
47	CTAATTAATCAT	0.162	621	10511	0.771	0.765	0.72
48	TGATTAATTGCT	0.252	867	15839	0.692	0.753	0.71
49	CCAATTAATCAT	0.102	368	6343	0.728	0.750	0.70
50	GCTATTAATCAT	0.157	550	9707	0.706	0.748	0.70

Supplementary Table S3, Slattery et al

	Exd-Lab	Exd-Pb	Exd-Dfd	Exd-Scr	Exd-Antp	Exd-Ubx	Exd-UbxIVa	Exd-AbdA	Exd-AbdB
1	GTAATCAATCAT (1.00)	GTAATCAATCAT (1.00)	ATGATTAATGAT (1.00)	ATGATTAATTGC (1.00)	GTAATAAATCAT (1.00)	ATGATTTATGAC (1.00)	GTAATAAATCAT (1.00)	GTAATAAATCAT (1.00)	GTCATAAATCAT (1.00)
2	ATGATTGATTAC (0.91)	ATCAGAACTCAT (0.96)	AATGATTAATGA (0.98)	GCAATTAATCAT (0.96)	ATGATTTATTAC (0.97)	GTCATAAATCAT (0.97)	GTCATAAATCAT (0.92)	ATGATTTATTAC (0.88)	AATGATTTATGA (0.97)
3	AATGATTGATTA (0.83)	AATGAGTTTTTA (0.93)	ATCATTAATCAT (0.97)	GTAATTAATCAT (0.94)	ATGATTTATTGC (0.90)	GTAATAAATCAT (0.97)	ATGATTTATTAC (0.92)	GTCATAAATCAT (0.84)	GGTCATAAATCA (0.96)
4	TAATCAATCATT (0.79)	ATCGATTTTTAA (0.90)	ATGATTAATGAC (0.95)	ATGATTAATTAC (0.91)	AATGATTTATTA (0.90)	ATGATTTATTAC (0.96)	ATGATTTATGAC (0.87)	ATGATTTATGAC (0.79)	ATGATTTATGAC (0.96)
5	GATGATTGATTA (0.72)	ATGGGTTTTTAA (0.89)	GATGATTAATGA (0.93)	GTCATTAATCAT (0.91)	GCAATAAATCAT (0.89)	AATGATTTATGA (0.90)	GCAATAAATCAT (0.86)	AATGATTTATTA (0.76)	GGTAATAAATCA (0.93)
6	AATGATTGATGA (0.70)	ATGATTGATTAC (0.89)	GTGATTAATGAT (0.92)	ATGATTAATGAC (0.91)	GGTAATAAATCA (0.88)	ATGATTTATTGC (0.85)	GGTAATAAATCA (0.81)	TAATAAATCATT (0.75)	GTAATAAATCAT (0.92)
7	GTCATCAATCAT (0.70)	AATGATTGATTA (0.88)	ATGATTAATTAT (0.92)	AATGATTAATGA (0.88)	TAATAAATCATT (0.80)	GATGATTTATGA (0.84)	ATGATTTATTGC (0.81)	GCAATAAATCAT (0.75)	ATGATTTATTAC (0.85)
8	ATGATTGATTAG (0.68)	ATTGATTTTTAA (0.82)	TGATTAATGATT (0.91)	ATGATTAATGGC (0.88)	GTCATAAATCAT (0.74)	TCATAAATCATT (0.83)	GGTCATAAATCA (0.80)	AATGATTTATGA (0.74)	AGTCATAAATCA (0.82)
9	ATGATTGATGAC (0.67)	TAATGAGTTTTT (0.82)	TGATTAATGATG (0.91)	AATGATTAATGG (0.86)	AATGATTTATGA (0.74)	ATGATTTATGAG (0.82)	AATGATTTATGA (0.77)	ATGATTTATTGC (0.73)	GCAATAAATCAT (0.82)
10	TAATCAATCATC (0.67)	TAATCAATCATT (0.80)	GTCATTAATCAT (0.90)	AATGATTAATTG (0.86)	TGATTTATTACC (0.72)	GCAATAAATCAT (0.81)	GCCATAAATCAT (0.75)	TCATAAATCATT (0.68)	GCCATAAATCAT (0.81)
11	ATGATTGATTGAG (0.67)	AATCGATTTTTA (0.78)	TCATTAATCATT (0.90)	GTAATCAATCAT (0.85)	ATGATTTATTGAC (0.72)	GGTCATAAATCA (0.81)	TAATAAATCATT (0.75)	GGTAATAAATCA (0.66)	TCATAAATCATT (0.80)
12	GGTAATCAATCA (0.65)	GGTAATCAATCA (0.76)	AATGATTAATTA (0.90)	GATGATTAATTG (0.85)	GATGATTTATTA (0.68)	AATGATTTATTA (0.80)	AATGATTTATTA (0.74)	GATGATTTATTA (0.65)	ATGATTTATGGC (0.77)
13	CTAATCAATCAT (0.65)	ATGTGTTTTTAA (0.74)	TGATTAATGATA (0.89)	ATGATTAATGAG (0.83)	AGTAATAAATCA (0.66)	GATGATTTATTA (0.80)	TGATTTATTACC (0.74)	TAATAAATCATC (0.63)	ATGATTTATTGC (0.77)
14	TCATCAATCATT (0.62)	AATTGATTTTTA (0.73)	TGATTAATGATC (0.89)	TAATTAATCATC (0.82)	AGCAATAAATCA (0.64)	ATGATTTATTGGC (0.80)	TGATTTATGACC (0.74)	AATGATTTATGG (0.62)	AATGATTTATTA (0.76)
15	CTCATCAATCAT (0.61)	AATGATTGATGA (0.73)	TATGATTAATGA (0.88)	TCATTAATCATT (0.82)	TCATAAATCATT (0.64)	GGTAATAAATCA (0.79)	TCATAAATCATT (0.73)	GATGATTTATGA (0.62)	AGTAATAAATCA (0.74)
16	GATGATTGATGA (0.58)	AATGGGTTTTTA (0.73)	AATCATTAATCA (0.88)	ATGATTGATTAC (0.82)	TAATAAATCATC (0.63)	ATGATTTATGGG (0.79)	TAATAAATCATC (0.73)	GCCATAAATCAT (0.62)	TGATTTATGACC (0.72)
17	AGTAATCAATCA (0.57)	GTTAATGATCAT (0.72)	ATGATTAATGAA (0.88)	GATGATTAATTA (0.82)	GCCATAAATCAT (0.62)	TAATAAATCATT (0.79)	GATGATTTATTA (0.73)	ATGATTTATGGC (0.62)	GCTATAAATCAT (0.72)
18	ATAATCAATCAT (0.57)	ATGGATTTTTAA (0.70)	TGATTAATGAT (0.88)	GATGATTAATGA (0.81)	GGTCATAAATCA (0.62)	TGATTTATGACC (0.79)	ATGATTTATGGG (0.72)	ATGATTTATGGG (0.61)	TGATTTATTACC (0.71)
19	ATGATTGATGGG (0.57)	ATTGGTTTTTAA (0.70)	ATAATTAATCAT (0.88)	AATGATTAATTA (0.81)	TGATTTATTACT (0.62)	ATGATTTATGAT (0.78)	GTCATAAATCAC (0.72)	AATGATTTATTG (0.59)	AATGATTTATGG (0.70)
20	TGATTGATTACC (0.56)	GTCATCAATCAT (0.68)	ATGATTAATGGT (0.87)	GATGATTAATGG (0.81)	ATGATTTATGGC (0.61)	TATAAATCATC (0.77)	ATGATTTATGGC (0.72)	GGTCATAAATCA (0.59)	TGATTTATGACT (0.70)
21	GCAATCAATCAT (0.55)	ACTAATGATCAT (0.67)	ATGATTAATTAC (0.86)	TAATTAATCATT (0.81)	TGATTTATTGCT (0.60)	AATGATTTATGG (0.76)	GATGATTTATGA (0.72)	CCATAAATCATT (0.58)	GTCGTAAATCAT (0.69)
22	CCCATCAATCAT (0.54)	ATGATTGATGAC (0.66)	TATCATTAATCA (0.85)	ATGATTTATTGC (0.80)	GGCATAAATCA (0.60)	TGATTTATTACC (0.75)	GTATAAATCAC (0.72)	CAATAAATCATT (0.58)	ATGATTTATAGC (0.69)
23	ATGATTGATTAT (0.54)	ATTAATGATCAT (0.66)	GATGATTAATTA (0.85)	ATGATTAATGGG (0.80)	AATGATTTATGG (0.60)	TCATAAATCATC (0.75)	GTGATTTATGAC (0.71)	ATGATTTATGAG (0.58)	ATCATAAATCAT (0.69)
24	TGATTGATTACT (0.53)	AATGTGTTTTTA (0.64)	CATCATTAATCA (0.85)	CAATTAATCATC (0.80)	AATGATTTATTG (0.57)	TATGATTTATGA (0.73)	ATGATTTATGAG (0.70)	ATAATAAATCAT (0.58)	GGTCGTAAATCA (0.68)
25	ATGATTGATCGA (0.53)	CATTAGTGCTAA (0.64)	GATCATTAATCA (0.85)	ATGATTGATTGC (0.79)	CAATAAATCATT (0.56)	ATCATAAATCAT (0.72)	AATGATTTATGG (0.70)	ATCATAAATCAT (0.57)	TAATAAATCATT (0.68)
26	TAATCAATCATA (0.53)	ATCGATATTTAA (0.64)	CATGATTAATGA (0.84)	CAATTAATCATT (0.79)	ATGATTTATGGG (0.55)	TGATTTATGACT (0.72)	ATGATTTATGAT (0.70)	ATGATTTATGAT (0.57)	ATGATTTATGAT (0.67)
27	TATGATTGATTA (0.51)	TTAAAAACTCAT (0.63)	AGTGATTAATGA (0.83)	ATGATTAATTAG (0.79)	ATAATAAATCAT (0.55)	GATGATTTATGG (0.72)	ATCATAAATCAT (0.69)	TGATTTATTACC (0.57)	ATGATTTACGAC (0.67)
28	ATGATTGATTGC (0.51)	AAATGATTGATT (0.63)	ATCATTAATCAC (0.83)	ATGATTAATTGG (0.78)	CCATAAATCATT (0.53)	CTCATAAATCAT (0.70)	CCCATAAATCAT (0.69)	TCATAAATCATC (0.57)	AATGATTTACGA (0.65)
29	ATGATTGATTGAT (0.51)	TAATTGATTTTT (0.63)	TCATTAATCATC (0.82)	ATGATTAATGAT (0.78)	TGATTTATGACC (0.52)	AGTCATAAATCA (0.69)	GTGATTTATTAC (0.69)	CCCATAAATCAT (0.57)	AGCAATAAATCA (0.65)
30	TCATCAATCATC (0.51)	AATCGATTATTA (0.63)	TGATTAATGACT (0.82)	GCCATTAATCAT (0.78)	ATGATTTATTAT (0.52)	GCCATAAATCAT (0.68)	GATGATTTATGG (0.69)	GATGATTTATGG (0.56)	GATGATTTATGA (0.65)
31	AATGATTGATGG (0.50)	AATGGATTTTTA (0.62)	GTAATTAATCAT (0.82)	TCATTAATCATC (0.78)	AAATGATTTATT (0.52)	TCATAAATCATA (0.68)	CCATAAATCATT (0.69)	AAATGATTTATG (0.56)	CCATAAATCATT (0.63)
32	AATGATTGATTG (0.50)	ATTAATGATCGT (0.62)	ATGATTAATGGA (0.81)	GAATGATTAATG (0.77)	CCCATAAATCAT (0.51)	AAATGATTTATG (0.68)	AGTCATAAATCA (0.68)	CTCATAAATCAT (0.55)	GAATGATTTATG (0.63)
33	ATCATCAATCAT (0.49)	CATTAACACTAA (0.62)	TAATTAATCATT (0.81)	GCAATCAATCAT (0.77)	GATGATTTATGA (0.51)	AATGATTTATTG (0.67)	TCATAAATCATC (0.68)	ATGATTTATTAT (0.54)	AAATGATTTATG (0.62)
34	GTAATCAATCAA (0.49)	TAATCGATTTTT (0.61)	ATGATTAATTAA (0.81)	ATGATTAATGGT (0.76)	GAATGATTTATT (0.50)	GAATGATTTATG (0.66)	AGTAATAAATCA (0.68)	TGATTTATGACC (0.54)	AGCCATAAATCA (0.62)
35	AATGATTGATCG (0.49)	AATCGATATTTA (0.61)	AGTCATTAATCA (0.81)	CCATTAATCATT (0.76)	AAATGATTTATG (0.50)	AGATGATTTATG (0.66)	TCATAAATCATA (0.68)	TATGATTTATGA (0.54)	GGCAATAAATCA (0.62)
36	TGATTGATCGAT (0.48)	TCATCAATCATT (0.60)	GTGATTAATGAC (0.80)	GTCATCAATCAT (0.76)	ATGATTTATGAG (0.50)	GTGATTTATGAC (0.65)	GGCATAAATCA (0.68)	TAATAAATCATA (0.53)	GGCCATAAATCA (0.61)
37	GGTCATCAATCA (0.48)	CATTAGTACTAA (0.60)	ATCATTAATCAA (0.79)	CTCATTAATCAT (0.76)	TGATTTATTGCC (0.50)	CCCATAAATCAT (0.65)	ATAATAAATCAT (0.68)	GTCGTAAATCAT (0.53)	AGTCGTAAATCA (0.61)
38	CAATCAATCATT (0.48)	AATTGGTTTTTA (0.59)	TTTGATTAATGA (0.79)	ATGATTTATTAC (0.75)	ATCATAAATCAT (0.50)	GATGATTTATTG (0.64)	TATGATTTATGA (0.68)	ATGATTTACGAC (0.53)	GGCTATAAATCA (0.60)
39	ATCGATCAATCA (0.48)	ATCGATTATTAA (0.59)	TCATTAATCATA (0.79)	ATGATTGATGAC (0.75)	ATGATTTATGAT (0.49)	AGTATAAATCAT (0.67)	CTCATAAATCAT (0.67)	GATGATTTATTG (0.53)	TGATTTATTACT (0.59)
40	TCGATCAATCAT (0.48)	CATTAAGTTAAT (0.59)	AATGATTAATGG (0.79)	GCAATAAATCAT (0.74)	AGTCATAAATCA (0.49)	CAATAAATCATT (0.63)	CAATAAATCATT (0.67)	TCATAATCATA (0.52)	AGCTATAAATCA (0.59)
41	CCATCAATCATT (0.47)	TAATGGGTTTTT (0.59)	TGATTAATGAAT (0.79)	GTAATAAATCAT (0.74)	GAATGATTTATG (0.48)	TGATTTATGACA (0.63)	CCATAAATCATC (0.67)	TATGATTTATTA (0.52)	GATGATTTATTA (0.58)
42	ATGATTGATTGG (0.47)	AAAATGATTGAT (0.59)	TGATTAATGACA (0.79)	ATCATTAATCAT (0.73)	AATAAATCATTT (0.47)	ATGATTTATGGT (0.62)	TGATTTATGACT (0.67)	CCATAAATCATC (0.52)	GATCATAAATCA (0.57)
43	ACTAATCAATCA (0.46)	CATTAACGTTAA (0.59)	TGATTAATGACC (0.79)	ATGATTAATTGT (0.73)	GATGATTTATGG (0.47)	TGATTTATTACT (0.62)	TAATAAATCATA (0.67)	CAATAAATCATC (0.52)	TCATAAATCATC (0.56)
44	AAATGATTGATT (0.46)	ATTAATGACCAT (0.58)	GGTCATTAATCA (0.79)	ATGATTAATAGC (0.73)	GTAATAAATCAC (0.47)	ATGATTTATTAT (0.62)	AGCATAAATCA (0.67)	GAATGATTTATG (0.52)	ATGATTTATGAG (0.56)
45	GTAATCAATCAC (0.46)	TTAAAAATCGAT (0.58)	TTGATTAATCAT (0.78)	CCCATTAATCAT (0.73)	CTCATAAATCAT (0.47)	CCATAAATCATT (0.62)	GTCGTAAATCAT (0.67)	AGTATAAATCA (0.52)	AGTGATTTATGA (0.56)
46	CCATCAATCATT (0.46)	CATTATACTAAT (0.57)	ACGATTAATCAT (0.78)	CCATTAATCATC (0.73)	TGCAATAAATCA (0.46)	GGATGATTTATG (0.61)	AATGATTTATTG (0.66)	CATAAATCATTT (0.51)	ATTGATTTATGA (0.56)
47	AAATGATTGATG (0.46)	AATCAATCATTT (0.57)	ATGATTGATGAT (0.78)	CTAATTAATCAT (0.72)	TATGATTTATTA (0.46)	TATGATTTATTA (0.61)	TGATTTATTACT (0.66)	GGCATAAATCA (0.51)	TGATTTATTGCT (0.55)
48	AATCAATCATTT (0.46)	GTTAATGATCGT (0.57)	ATGATTAATGAG (0.78)	TGATTAATTGCT (0.71)	TCATAAATCATC (0.46)	AGCATAAATCA (0.61)	CAATAAATCATC (0.66)	AAATGATTTATT (0.51)	CTCATAAATCAT (0.55)
49	CGATCAATCATT (0.46)	CATTAGACTAAT (0.57)	ATTGATTAATGA (0.77)	CCAATTAATCAT (0.70)	GTGATTTATTAC (0.46)	TGATTTATTGCT (0.61)	GTCATAAATCAA (0.66)	AATAAATCATTT (0.50)	GATGATTTATGG (0.55)
50	GATGATTGATGG (0.45)	CATTAGTCTAAT (0.57)	AAGATTAATGAT (0.77)	GCTATTAATCAT (0.70)	AATAAATCATTC (0.46)	TAATAAATCATA (0.60)	ATGATTTATTAT (0.66)	AGCAATAAATCA (0.49)	TGATTTACGACC (0.55)

Supplementary Table S4, Slattery et al

	Exd-Lab Round 3	**Exd-Pb** Round 3	**Exd-Dfd** Round 3	**Exd-Scr** Round 4	**Exd-Antp** Round 3	**Exd-UbxIa** Round 4	**Exd-UbxIVa** Round 3	**Exd-AbdA** Round 3	**Exd-AbdB** Round 3	**Exd-Exd** Round 1
Exd-Hox										
TGATTTAT	0.61%	0.11%	2.60%	10.19%	48.15%	66.67%	65.54%	61.63%	40.36%	0.12%
AGATTTAT	0.03%	0.02%	0.06%	0.21%	0.46%	0.81%	3.02%	0.29%	0.95%	0.07%
TAATTTAT	0.05%	0.50%	0.36%	0.70%	0.47%	3.21%	9.59%	0.99%	1.07%	0.07%
TGATTAAT	2.42%	0.96%[1]	66.30%	59.61%	1.65%	1.13%	2.03%	2.79%	0.89%	0.09%
AGATTAAT	0.10%	0.77%[1]	2.40%	1.22%	0.07%	0.12%	0.09%	0.03%	0.08%	0.06%
TAATTAAT	0.99%	7.03%[1]	22.49%	22.90%	0.74%	0.92%	0.68%	0.79%	0.47%	0.07%
TGATTGAT	80.06%	0.36%	14.99%	11.01%	10.52%	0.48%	2.56%	9.22%	8.31%	0.37%
AGATTGAT	1.87%	0.03%	0.45%	0.37%	0.42%	0.14%	0.21%	0.26%	0.57%	0.11%
TAATTGAT	0.88%	1.43%	1.91%	1.24%	0.36%	0.23%	0.28%	0.29%	0.38%	0.12%
TGATTTAC	0.11%	0.06%	0.24%	0.11%	0.63%	2.59%	4.79%	4.52%	4.97%	0.11%
AGATTTAC	0.01%	0.01%	0.01%	0.01%	0.02%	0.04%	0.21%	0.03%	0.15%	0.06%
TAATTTAC	0.01%	1.28%	0.06%	0.13%	0.03%	0.25%	0.81%	0.08%	0.19%	0.08%
TGACAAAT	0.02%	0.03%	0.12%	0.08%	0.11%	0.02%	0.02%	0.04%	0.11%	0.08%
AGACAAAT	0.01%	0.07%	0.01%	0.02%	0.02%	0.01%	0.01%	0.01%	0.02%	0.05%
TAACAAAT	0.03%	0.09%	0.02%	0.06%	0.05%	0.09%	0.05%	0.03%	0.05%	0.08%
TGATGGAT	3.61%	0.02%	0.67%	0.17%	1.00%	0.03%	0.10%	1.13%	0.82%	0.15%
AGATGGAT	0.06%	0.00%	0.01%	0.01%	0.02%	0.02%	0.00%	0.01%	0.02%	0.05%
TAATGGAT	0.04%	0.22%	1.45%	0.30%	0.03%	0.03%	0.03%	0.03%	0.03%	0.05%
TGATAAAT	0.08%	0.07%	1.44%	1.12%	0.21%	0.44%	0.62%	0.33%	0.26%	0.07%
AGATAAAT	0.02%	0.10%	0.06%	0.10%	0.09%	0.17%	0.37%	0.14%	0.08%	0.06%
TAATAAAT	0.22%	0.23%	0.84%	4.26%	17.80%[2]	22.85%[2]	19.34%[2]	18.77%[2]	9.33%[2]	0.08%
TGACTAAT	0.10%	0.69%	0.24%	0.22%	0.06%	0.05%	0.05%	0.05%	0.07%	0.05%
AGACTAAT	0.15%	0.95%	0.04%	0.11%	0.04%	0.06%	0.04%	0.04%	0.03%	0.04%
TAACTAAT	0.16%	1.25%	0.08%	0.19%	0.06%	0.10%	0.06%	0.05%	0.04%	0.07%
TGATTAAC	0.63%	0.59%	0.50%	0.29%	0.09%	0.07%	0.10%	0.16%	0.13%	0.10%
AGATTAAC	0.01%	0.13%	0.01%	0.02%	0.01%	0.02%	0.01%	0.01%	0.02%	0.07%
TAATTAAC	0.05%	12.16%[1]	1.16%	0.83%	0.09%	0.14%	0.06%	0.03%	0.10%	0.09%
TGATCGAT	4.77%	0.04%	0.21%	0.07%	0.18%	0.01%	0.03%	0.17%	0.16%	0.06%
AGATCGAT	0.19%	0.01%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%	0.02%	0.03%
TAATCGAT	1.29%	0.40%	0.32%	0.20%	0.02%	0.02%	0.02%	0.03%	0.02%	0.04%
Exd-Exd										
TGA[T|C].{5,6}TGA[T|C]	1.04%	0.51%	1.57%	0.62%	6.30%	0.87%	1.01%	2.24%	5.35%	0.42%
TGA[T|C].{5,6}[A|G]TCA	1.26%	0.12%	2.62%	0.80%	4.73%	0.72%	0.92%	1.41%	4.08%	0.20%
[A|G]TCA.{5,6}TGA[T|C]	0.52%	0.21%	0.09%	0.09%	0.45%	0.03%	0.03%	0.17%	0.44%	0.22%
Hox-Hox										
TAATTA.{1,4}TAATTA	0.00%	77.46%	0.05%	0.04%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%
Any of the above	90.83%	31.87%	94.08%	86.68%	72.30%	77.08%	89.94%	82.94%	64.96%	3.22%

[1] These motifs overlap considerably with the Hox-Hox homodimer motif, and therefore have inflated percentages

[2] The reverse complement of these motifs overlaps with the red motifs, which are preferred by the Class 3 proteins