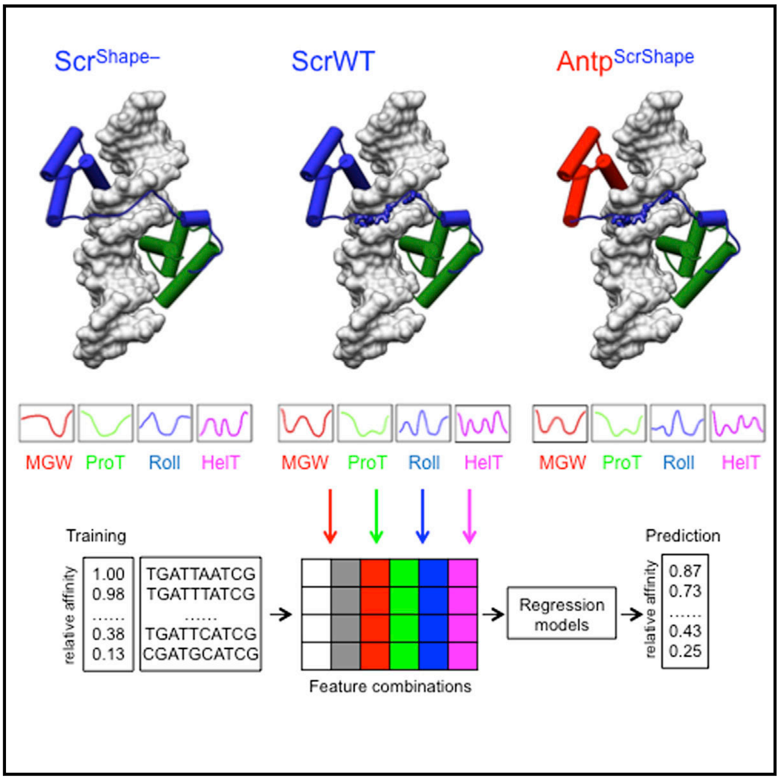
Cell

Article

Deconvolving the Recognition of DNA Shape from Sequence

Graphical Abstract



Authors

Namiko Abe, Iris Dror, ..., Remo Rohs, Richard S. Mann

Correspondence

rohs@usc.edu (R.R.),
rsm10@columbia.edu (R.S.M.)

In Brief

DNA structural features play an independent and direct role in binding specificity by Hox proteins, and knowledge of these features facilitates the de novo prediction of DNA binding specificities.

Highlights

- DNA shape-recognizing residues play a direct role in Hox-DNA binding specificity
- These residues are sufficient to swap the specificity of one Hox protein to another
- Accuracy of binding specificity predictions improves by including DNA shape features
- Machine learning reveals positions in the binding site where DNA shape is recognized

Accession Numbers

GSE65073

Abe et al., 2015, Cell *161*, 307–318
April 9, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.02.008

Supplemental Information

EXTENDED EXPERIMENTAL PROCEDURES

High-Throughput DNA Shape Prediction

All sequences selected in R3 of SELEX with a count of at least 25 were aligned based on the TGAYNNAY (Exd-Hox heterodimers) or TAAT (Hox monomers) motifs, where N can be any nucleotide and Y represents C or T. Sequences with multiple occurrences of these motifs were removed from the analysis. Four DNA structural features were derived for these sequences (numbers of sequences for each Hox variant are listed in Table S2 and S3) from a high-throughput DNA shape prediction method, which is based on mining DNA structural information from a pentamer library derived from all-atom Monte Carlo simulations (Zhou et al., 2013). For each nucleotide position of the aligned sequences, MG width and propeller twist (ProT) were predicted. Roll and helix twist (HelT) were predicted using the same approach for each base pair step.

Euclidean Distance Comparisons

The average MG width at each position of all sequences selected by Exd-ScrWT and Exd-AntpWT was calculated and defined as reference ScrWT and AntpWT shape preference, respectively. Next, MG width profiles were calculated for each of the sequences selected by Exd-ScrWT, Exd-Scr mutants, Exd-AntpWT, and Exd-Antp mutants with a relative affinity > 0.8 and compared to the reference ScrWT shape reference using Euclidean distances, where a low Euclidean distance score implied high similarity between the MG width pattern of the given sequence and the reference ScrWT shape preference. An analogous approach was used to score the shape preference similarity to AntpWT binding, in which the Euclidean distance was calculated between the MG width of each sequence and the reference AntpWT shape preference. This analysis was based on 16-mers because MG width is not defined for two nucleotides at each end, resulting in a MG width pattern for the central 12-mers.

L2-Regularized Multiple Linear Regression

To predict the relative binding affinity for each of the sequences bound by any of the Hox monomers, Exd-Hox heterodimers, and Exd-Hox mutant heterodimers, we trained L2-regularized multiple linear regression (MLR) models (Yang et al., 2014). To measure the predictive power of the models, a 10-fold cross-validation was performed with an embedded 10-fold cross-validation on the training set to determine the optimal λ parameter. The source code for the DNA shape prediction and feature mapping is available for download at: http://rohslab.cmb.usc.edu/Cell2015/

Regression Models for Predicting Binding Specificities Quantitatively

We trained different categories of models that (i) encoded the nucleotide sequence of each of the bound sequences as binary features (sequence models), (ii) encoded different combinations of the DNA shape features MG width, ProT, Roll, and HelT (shape models), and (iii) combined nucleotide sequence and DNA shape features at the corresponding position (sequence+shape models). To measure the predictive power of each of the models, we calculated the coefficient of determination R^2 between the predicted and experimentally determined logarithm of relative binding affinities using 10-fold cross validation. We used all 14-mer sequences from R3 of the selection with a count of > 50, aligned based on the TGAYNNAY core motif for heterodimers, and the logarithm of the relative binding affinity as response variable. We used 14-mers in this analysis as a trade-off between sequence length and read coverage. Sequences with the core motif not located in the center resulted in missing flanks due to the alignment. We assigned features with a value of zero to these end positions. Sequences that did not contain the motif or contain more than one core motif were not included in the analysis.

As a form of feature selection, we trained variants of these models where we added or removed features at specific positions and evaluated the performance of these models based on a ΔR^2 with respect to a reference model. Shape features at position *i* include MG width and ProT at position *i*, whereas the definition for Roll and HelT includes the base pair steps between nucleotides *i-1* and *i* as well as *i* and *i+1* (Zhou et al., 2013). For the analysis of monomer binding specificities, we aligned all single occurrences of TAAT motifs in 9-mer sequences.

To evaluate the robustness of our results, we compared MLR-based models with models trained using support vector regression (ε-SVR) with a linear kernel (Gordân et al., 2013; Zhou et al., 2015) by calculating the Pearson correlation between R^2s derived from the two methods. We performed the ε-SVR analysis for sequence and sequence+shape models based on 10-fold cross validation for Exd-Hox WTs using 16-mer relative binding affinities as response variable and determined the hyper-parameters *C* and ε in a grid search using nested cross validation.

Classification Models for Distinguishing ScrWT-like and AntpWT-like Binding Specificities

To use sequence and shape features to classify Hox binding specificities, we aligned 14-mers selected by Exd-ScrWT or Exd-AntpWT according to the core motif TGAYNNAY. For the 14-mers with a single occurrence of the core motif, the top 50% with the highest relative affinities were selected from the datasets prepared as described above as the preferred binding sites for ScrWT, AntpWT, and the mutants. As a result, the ScrWT dataset comprised 7078 and the AntpWT dataset 4962 sequences. Among these sequences, we removed 2416 sequences that were shared between both datasets, resulting in 4662 ScrWT preferred sites (assigned the label +1) and 2546 AntpWT preferred sites (assigned the label –1). The models were evaluated based on this training data using L2-regularized MLR and 10-fold cross-validation, and area under the receiver-operating characteristic curve (AUC) was used as performance measure.

The trained models were used to classify the top 50% aligned binding sites preferred by the mutants. The MLR prediction resulted in a continuous number, which was converted into a binary classification measure based on whether the response variable is >0 for ScrWT-like or <0 for AntpWT-like binding specificities. The Pearson correlation was calculated between the response variable (class labels) and the normalized MG width at each position to reveal which positions affect the classification most (positions with strong correlation, either positive or negative). We normalized MG width by dividing the difference of MG width at position *i* and the MG width mean over all unique pentamers by the standard deviation in MG width over all 512 possible pentamers as derived from the DNA-shape method (Zhou et al., 2013). All shape parameters were normalized by the same scheme for the aforementioned MLR analysis.

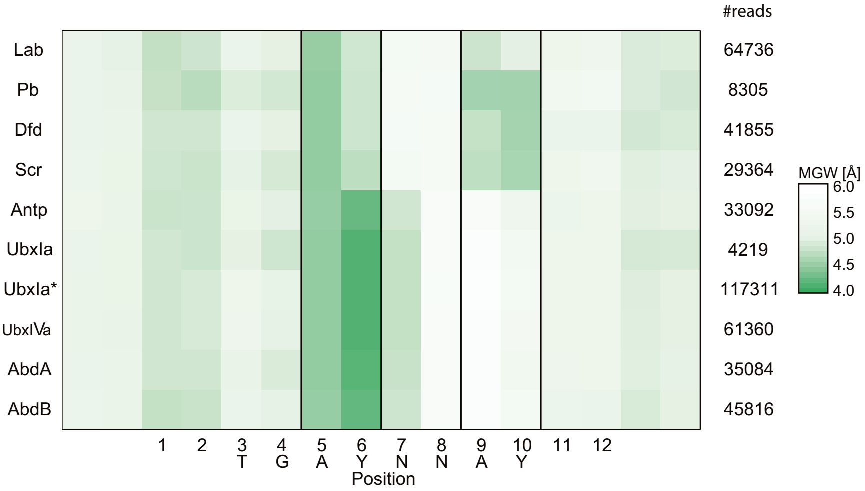


Figure S1. Anterior and Posterior Hox Proteins Select for Sequences with Distinct Minor Groove Shapes, Related to Figure 1

Heat map of the average MGW at each position of 16-mers selected by each Exd-HoxWT heterodimer. Except for Pb and UbxIa*, the SELEX data from Slattery et al. (2011) were re-analyzed using a common error cutoff of 20%. UbxIa* represents new SELEX data due to low counts in the previous dataset. In addition, because the previous dataset used a truncated form of Pb, we carried out a new SELEX experiment with full-length Pb. Dark green represents narrow minor grooves while white represents wider minor grooves. The numbers to the right of the heat map indicate the number of sequences analyzed for each complex. Black lines demarcate where Arg5 inserts into the minor groove (A_5Y_6) and, for Scr, where Arg3 and His-12 insert into the minor groove (A_9Y_{10}).

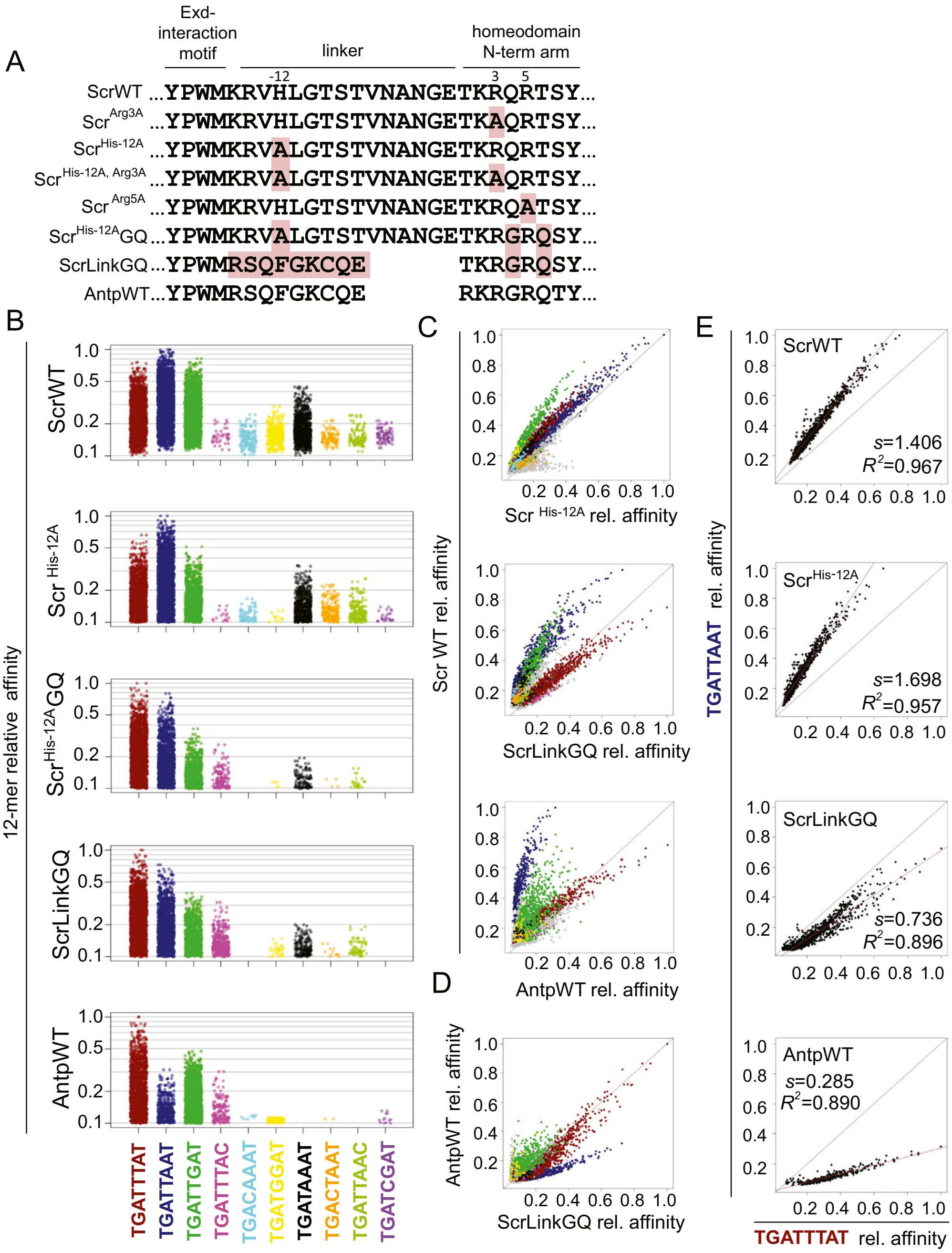


Figure S2. Binding Specificities of Scr Variants with Antp's N-Terminal Arm and Linker Sequences, Related to Figure 1

(A) Amino acid sequences of all Scr variants. Numbering is relative to the first residue in the homeodomain. Only sequences from the Exd-interaction motif YPWM through the homeodomain N-terminal arm are shown. The rest of the protein is wild-type in all cases. Red highlights the modified residues.

(B) 12-mer relative affinities of Scr variants.

(C) Comparative specificity plots of sequences selected by Exd-ScrWT versus those selected by Exd-Scr variants. Each point represents a unique 12-mer that is color-coded according to the core 8-mer it contains. Gray points represent 12-mers that do not contain any of the ten core 8-mers. The black line indicates $y = x$. Comparative specificity plot of Exd-ScrWT versus Exd-AntpWT is shown at the bottom.

(D) Comparative specificity plot of Exd-AntpWT versus Exd-ScrLinkGQ 12-mer affinities showing that the binding site preferences of ScrLinkGQ are distinct from those of AntpWT.

(E) Plots comparing the relative affinities of blue motifs (TGATTAAT) (y axis) to red motifs (TGATTTAT) (x axis). Black line indicates $y = x$, and the red line plots a linear regression trend line. The slope of the trend line and coefficient of determination R^2 of the data are indicated.

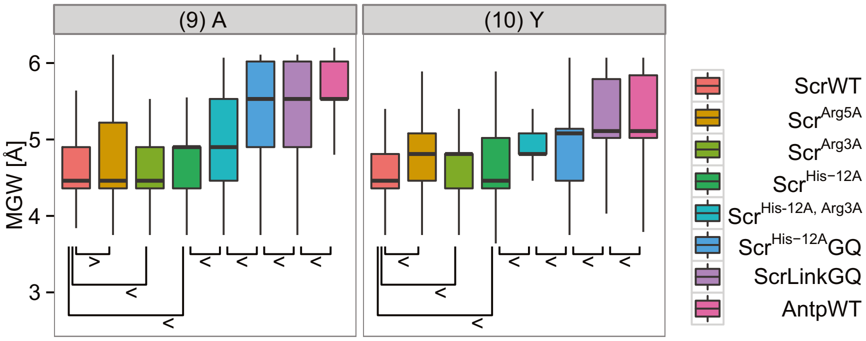


Figure S3. Shape Readout Properties of Scr Variants with Antp's N-Terminal Arm and Linker Sequences, Related to Figure 1

Box and whisker plots represent the distribution of the MGW at positions A_9 and Y_{10} of all 16-mer sequences selected by each Exd-Hox heterodimer. Wilcoxon test p values between dataset pairs (indicated by brackets) were calculated. The direction of the arrow indicates statistical significance of dataset in one direction (< indicates that the values of the right dataset are larger than the left dataset with $p < 0.001$; > indicates that the values of the right dataset are smaller than the left dataset with $p < 0.001$).

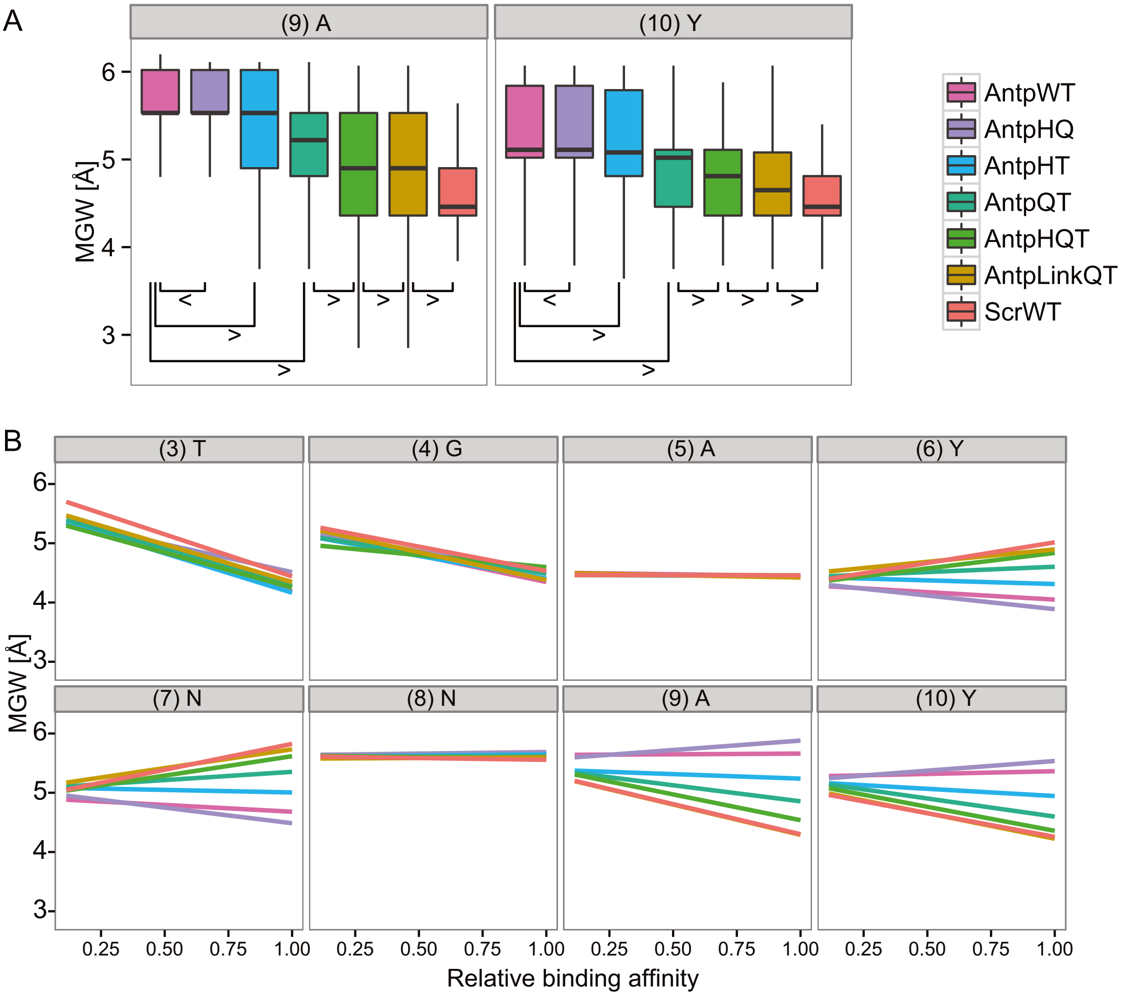


Figure S4. Q4, T6, and Linker Residues Contribute to Narrow Minor Groove Recognition, Related to Figure 4

(A) Box and whisker plots represent the distribution of the MGW at positions A_9 and Y_{10} of all 16-mer sequences selected by each Exd-Hox heterodimer. Wilcoxon test p values between pairs of datasets (indicated by brackets) were calculated. The direction of the arrow indicates statistical significance of dataset in one direction (< indicates that the values of the right dataset are larger than the left dataset with $p < 0.001$; > indicates that the values of the right dataset are smaller than the left dataset with $p < 0.001$).

(B) Linear regression trend lines representing the relationship between SELEX relative affinities and the MGW at positions 3 to 10 within the 16-mer. Although the raw data are fairly scattered, it is noteworthy that differences in these trend lines are only observed at four of these eight positions.

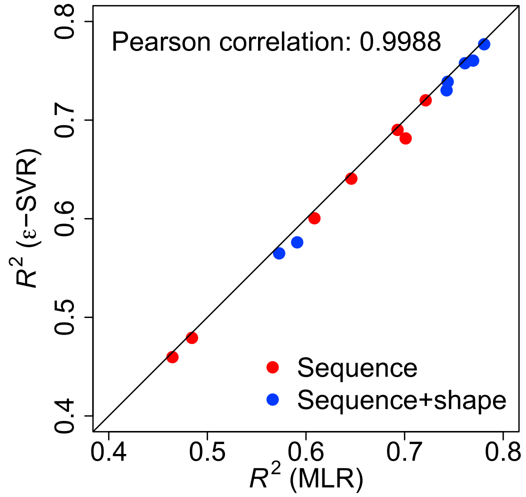


Figure S5. Comparison of Model Evaluation Based on Support Vector Regression versus Multiple Linear Regression, Related to Figure 6

The coefficients of determination R^2 derived from support vector regression (ε-SVR; y axis) and L2-regularized multiple linear regression (MLR; x axis) for sequence models (red) and sequence+shape models (blue) for 7 Exd-Hox WTs, together, result in a Pearson correlation close to 1. The heterodimer datasets for Lab and UbxIVa were too large for the ε-SVR analysis due to the required grid search to determine the *C* and ε hyper-parameters. The comparison of the two machine learning methods used 16-mer relative binding affinity data, as shown in Figures 2 and 4A.

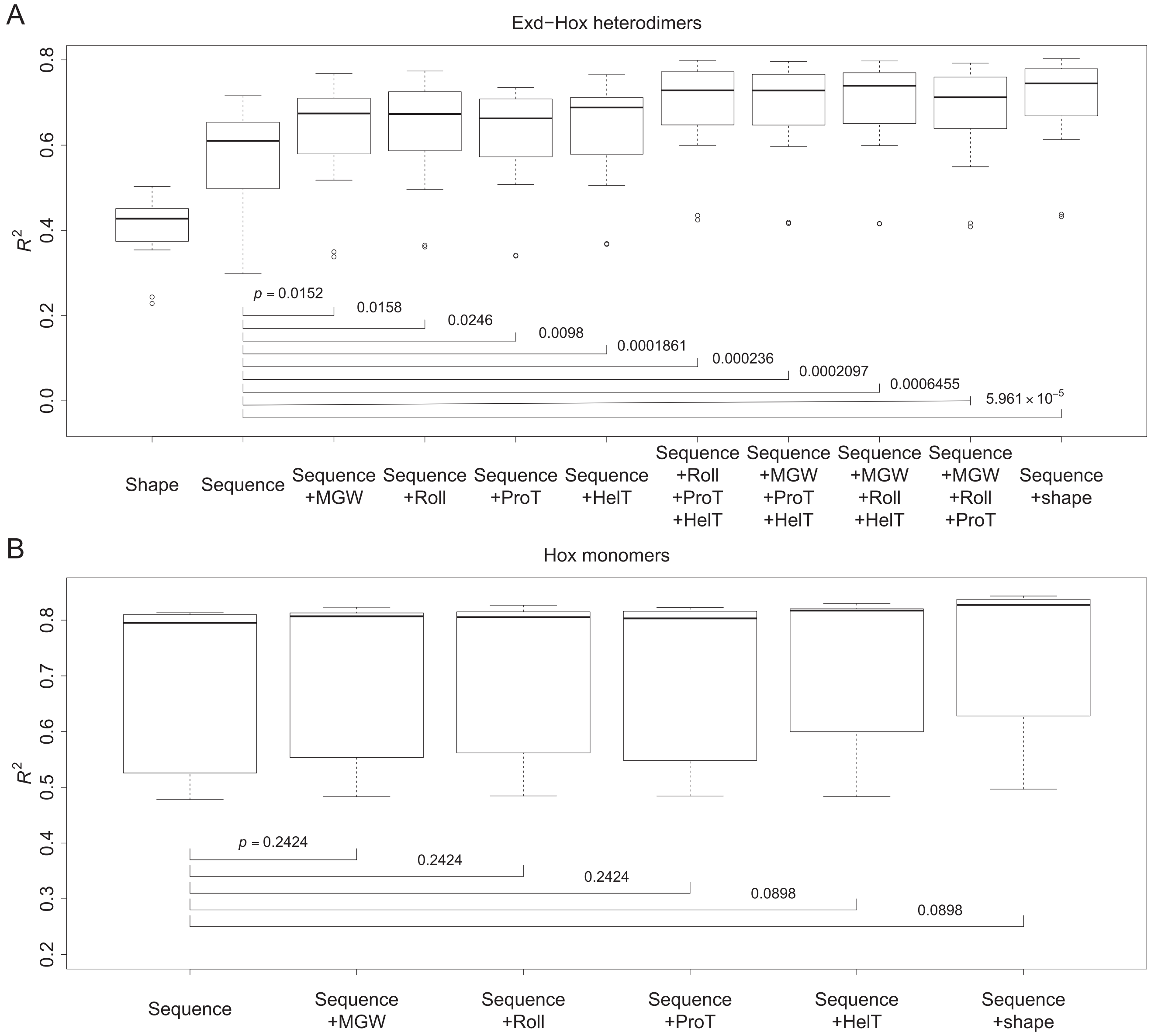


Figure S6. DNA Shape Features Improve Quantitative Predictions of DNA Binding Specificities of Exd-Hox Heterodimers and Hox Monomers, Related to Figure 6

(A) Adding one of the four shape features (MGW, Roll, ProT and HelT) to a sequence model improves DNA binding specificity predictions of Exd-Hox heterodimers about equally well, measured based on the coefficient of determination R^2, whereas addition of any three shape features simultaneously results in a compounded effect and addition of all four shape features at the same time results in the largest effect. *P*-values calculated based on a one-sided Mann-Whitney test for the sequence model versus shape-augmented models demonstrate the significance of the effect, with the sequence+shape model resulting in the most significant improvement. The centerline of the box plots represents the median, the edge of the box the 1st and 3rd quartile, and the whiskers indicate minimum/maximum values within 1.5 times the interquartile from the box.

(B) Adding one of the four shape features (MGW, Roll, ProT and HelT) or all four shape features at once to a sequence model has only a modest effect on DNA binding specificity predictions of Hox monomers, measured based on the coefficient of determination R^2 and one-sided Mann-Whitney test p values. While this observation might in part be due to the stringent filtering of only TAAT motifs, it demonstrates the larger role of DNA shape on Exd-Hox heterodimer binding. The box plots are defined in (A).

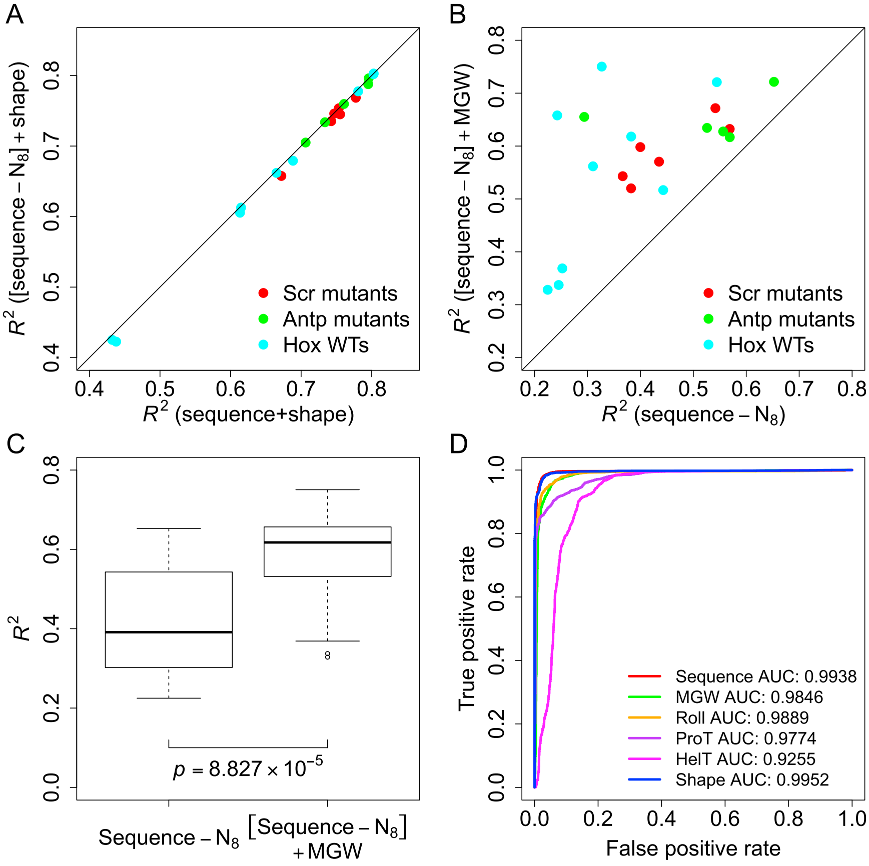


Figure S7. Models that Deconvolve DNA Sequence and Shape Further Demonstrate the Additional Information Contained by Shape-Based Models, Related to Figure 7

(A) Scatter plot representing the coefficient of determination R^2 obtained using a sequence+shape model (x axis) compared to a sequence+shape model with sequence information removed at the N_8 position ([sequence–N_8]+shape model) (y axis). Removing sequence features at the N_8 position where sequence is most variable across the selected sequences has essentially no effect on model accuracy as all points lie on or close to the diagonal. Each point represents a different Hox variant and is color-coded as indicated.

(B) Scatter plot representing the coefficient of determination R^2 obtained using a sequence–N_8 model (x axis) compared to the sequence–N_8 model with MG width (MGW) features added to all positions (y axis). All points above the diagonal line represent complexes in which the MGW-augmented model improves the prediction accuracy of the logarithm of relative binding affinities in comparison to a model that does not use this information. The color code for different heterodimers is equivalent to (A). Removing sequence features at the N_8 position deconvolves the contribution of MGW to model accuracy from sequence.

(C) Box plots illustrate that adding MGW to the sequence–N_8 model improves DNA binding specificity predictions of Exd-Hox heterodimers, measured based on the coefficient of determination R^2. One-sided Mann-Whitney p values demonstrate the significance of the effect. The centerline of the box plots represents the median, the edge of the box the 1st and 3rd quartile, and the whiskers indicate minimum/maximum values within 1.5 times the interquartile from the box.

(D) Classification models based on individual shape features perform well in distinguishing AntpWT-like from ScrWT-like binding specificities, measured based on area under the-receiver-operating characteristic curve (AUC), with the combined shape model performing equally well as the sequence model.

Table S1. Oligonucleotide Sequences, Related to Experimental Procedures

Table of oligonucleotide sequences used for SELEX experiments.

Oligo name	Sequence
16mer Multiplex 1*	GTTCAGAGTTCTACAGTCCGACGATCTGG[N_{16}]CCA**GCTG**TCGTATGCCGTCTTCTGCTTG
16mer Multiplex 2*	GTTCAGAGTTCTACAGTCCGACGATCTGG[N_{16}]CCA**CGTC**TCGTATGCCGTCTTCTGCTTG
16mer Multiplex 3*	GTTCAGAGTTCTACAGTCCGACGATCTGG[N_{16}]CCA**GAAC**TCGTATGCCGTCTTCTGCTTG
16mer Multiplex 4*	GTTCAGAGTTCTACAGTCCGACGATCTGG[N_{16}]CCA**AGAG**TCGTATGCCGTCTTCTGCTTG
16mer Multiplex 5*	GTTCAGAGTTCTACAGTCCGACGATCTGG[N_{16}]CCA**ACCT**TCGTATGCCGTCTTCTGCTTG
Hox-Exd tracking	GCTATACTGTGCTATCCACAGTTCAGAGTCGAAAATGATTGATTACCGCTGGTCACTGGTCGTTTCCCTCTT

*Barcodes are in bold. [N_{16}] represents the 16 randomized bases

Table S2. Number of Sequences Included in High-Throughput DNA Shape Analysis (Exd-Hox Heterodimers, Related to Experimental Procedures

Middle column displays total number of sequences in R3 with counts ≥ 25. Right column displays number of sequences that contained TGAYNNAY motif and were further used in the shape analysis. This table refers to the sequences used in the heat maps in Figures 2, 4, and S1.

Hox+Exd	Total number of 16mers	Total number of 16mers with TGAYNNAY
ScrWT	30872	29364
ScrArg5A	211883	197562
ScrArg3A	15602	14833
ScrHis-12A, Arg3A	31231	29135
ScrHis-12A	133860	93380
ScrHAGQ	119198	114164
ScrLinkGQ	186326	165253
AntpWT	39585	33092
AntpHQ	138336	133623
AntpHT	97125	90094
AntpQT	74707	72733
AntpHQT	19005	18237
AntpLinkQT	121451	116566
Lab	78326	64736
Pb	16218	8305
Dfd	48432	41855
UbxIa	4464	4219
UbxIa*	132537	117311
UbxIVa	65604	61360
AbdA	37068	35084
AbdB	55492	45816

Table S3. Number of Sequences Included in High-Throughput DNA Shape Analysis (Hox Monomers, Related to Experimental Procedures

Middle column displays total number of sequences in R3 with counts ≥ 25. Right column displays number of sequences that contained TAAT (Hox monomers) and were further used in the shape analysis. This table refers to the sequences used in Figure S6.

Hox monomer	Total number of 9mers	Total number of 9mers with TAAT
AbdB	19645	3048
Dfd	4717	3092
Lab	130234	5520
Pb	2706	2152
Scr	38034	5356
UbxIVa	9818	4366