Cell

# Deconvolving the Recognition of DNA Shape from Sequence

Namiko Abe,[1,2] Iris Dror,[3,7] Lin Yang,[3] Matthew Slattery,[8] Tianyin Zhou,[3] Harmen J. Bussemaker,[9] Remo Rohs,[3,4,5,6,*] and Richard S. Mann[1,2,*]
[1]Department of Biochemistry and Molecular Biophysics
[2]Department of Systems Biology
Columbia University, New York, NY 10032, USA
[3]Molecular and Computational Biology Program, Department of Biological Sciences
[4]Department of Chemistry
[5]Department of Physics and Astronomy
[6]Department of Computer Science
University of Southern California, Los Angeles, CA 90089, USA
[7]Department of Biology, Technion – Israel Institute of Technology, Haifa 32000, Israel
[8]Department of Biomedical Sciences, University of Minnesota Medical School, Duluth, MN 55812, USA
[9]Department of Biological Sciences, Columbia University, New York, NY 10032, USA
*Correspondence: rohs@usc.edu (R.R.), rsm10@columbia.edu (R.S.M.)
http://dx.doi.org/10.1016/j.cell.2015.02.008

## SUMMARY

**Protein-DNA binding is mediated by the recognition of the chemical signatures of the DNA bases and the 3D shape of the DNA molecule. Because DNA shape is a consequence of sequence, it is difficult to dissociate these modes of recognition. Here, we tease them apart in the context of Hox-DNA binding by mutating residues that, in a co-crystal structure, only recognize DNA shape. Complexes made with these mutants lose the preference to bind sequences with specific DNA shape features. Introducing shape-recognizing residues from one Hox protein to another swapped binding specificities in vitro and gene regulation in vivo. Statistical machine learning revealed that the accuracy of binding specificity predictions improves by adding shape features to a model that only depends on sequence, and feature selection identified shape features important for recognition. Thus, shape readout is a direct and independent component of binding site selection by Hox proteins.**

## INTRODUCTION

Precise control of gene expression relies on the ability of transcription factors to recognize specific DNA binding sites. Two distinct modes of protein-DNA recognition have been described: base readout, the formation of hydrogen bonds or hydrophobic contacts with functional groups of the DNA bases, primarily in the major groove (Seeman et al., 1976), and shape readout, the recognition of the 3D structure of the DNA double helix (Rohs et al., 2009a). The importance of shape readout has been inferred from crystal structures of protein-DNA complexes (Joshi et al., 2007; Kitayner et al., 2010; Meijsing et al., 2009; Rohs et al., 2009b) and from structural features of DNAs selected by DNA-binding proteins in high-throughput binding assays (Dror et al., 2014; Gordân et al., 2013; Lazarovici et al., 2013; Slattery et al., 2011; Yang et al., 2014). However, as DNA shape is a function of the nucleotide sequence, it is difficult to tease apart whether a DNA binding protein favors a particular binding site because it recognizes its nucleotide sequence or, alternatively, structural features of the DNA molecule. Thus, whether DNA shape is a direct determinant of protein-DNA recognition remains an open question. In addition to being a potentially important mode of DNA recognition, if DNA binding proteins directly use shape readout then incorporating DNA structural information should significantly improve models for predicting DNA binding specificity, which remains challenging with existing methods (Slattery et al., 2014; Weirauch et al., 2013).

We previously described a role for DNA shape in the recognition of specific binding sites by the Hox family of transcription factors, which in vertebrates and *Drosophila* specify the unique characteristics of embryonic segments along the anterior-posterior axis (Joshi et al., 2007; Mann et al., 2009; Slattery et al., 2011). Using in vitro selection combined with deep sequencing (SELEX-seq), which examines millions of sequences in an unbiased manner, we found that while Hox proteins bind highly similar sequences as monomers, heterodimerization with the cofactor Extradenticle (Exd) uncovers latent DNA binding specificities (Slattery et al., 2011). High-throughput DNA shape predictions (Zhou et al., 2013) for sequences selected by each Exd-Hox complex (containing the motif NGAYNNAY) revealed that anterior and posterior Hox proteins prefer sequences with distinct minor groove (MG) topographies. Whereas all Exd-Hox complexes preferred sequences with a narrow MG near the AY of the Exd half-site (NGAY), only anterior Hox proteins (Lab, Pb, Dfd, and Scr) selected for sequences containing an additional minimum in MG width at the AY of the Hox half-site (NNAY) (Figures 1A and S1) (Slattery et al., 2011). However, this study, as well as analyses of other protein-DNA complexes (Gordân et al., 2013; Yang et al., 2014), did not rule out the

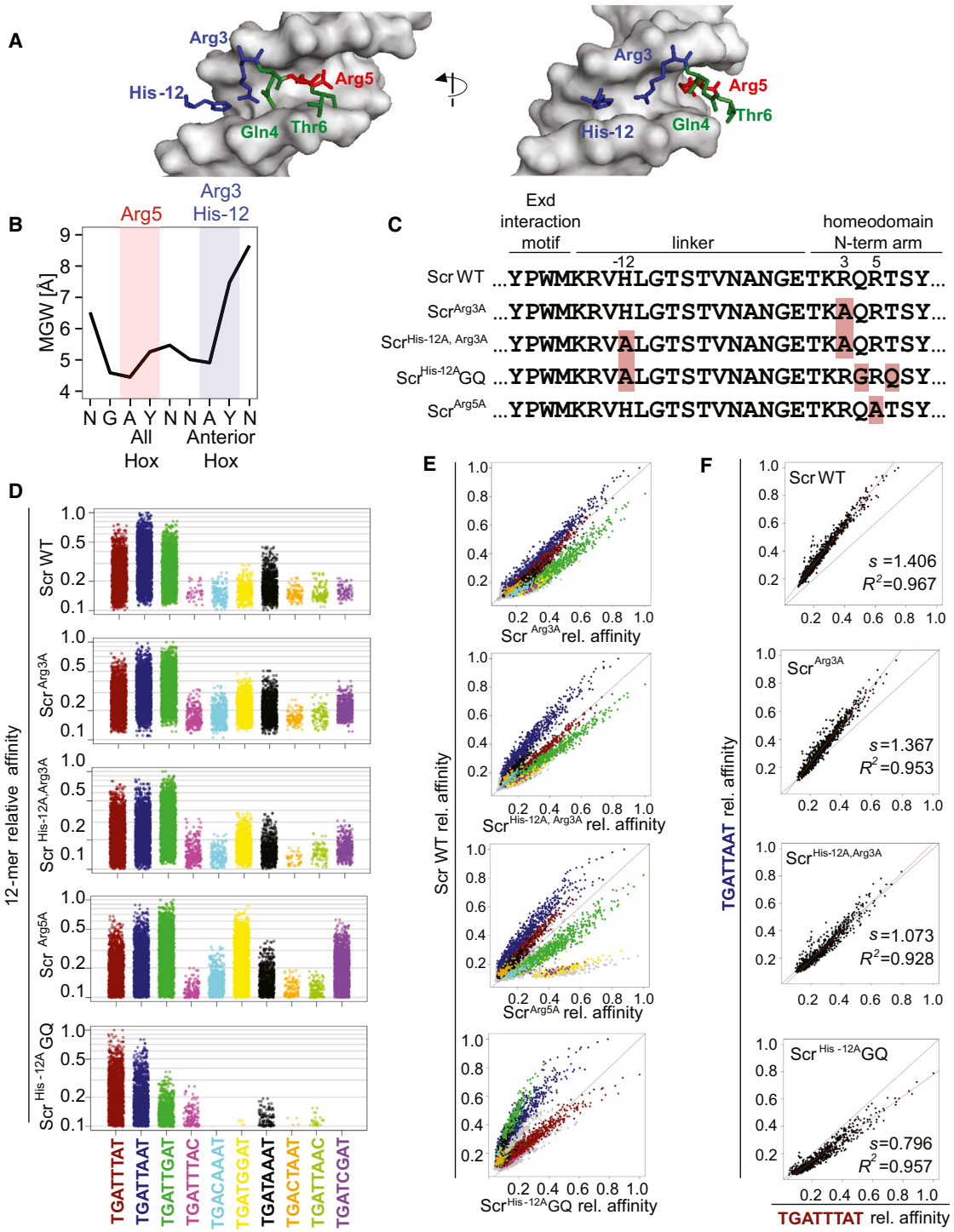


**Figure 1. Scr's Narrow-MG Recognizing Residues Are Required for Binding Specificity**

(A) Two views of the Exd-Scr heterodimer bound to the Scr-specific target *fkh250* (Protein Data Bank [PDB] ID 2R5Z) (Joshi et al., 2007).

(B) Plot of MG width derived from the Exd-Scr co-crystal structure showing that Arg5 (red) inserts into the MG width minimum at the Exd half-site (NGAY) while Arg3 and His-12 (blue) insert into the MG width minimum at the Hox half-site (NNAY).

(C) Amino acid sequences of 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 mutated residues.

(D) 12-mer relative affinities of binding sites selected by each Scr variant in complex with Exd are color-coded according to the ten most frequently observed Exd-Hox binding sites.

*(legend continued on next page)*

possibility that these shape preferences were merely a secondary consequence of base readout preferences.

A key prediction of the shape-recognition model is that if the residues that recognize a distinct structural feature of the DNA, for example a local minimum in MG width, are mutated then the transcription factor should no longer prefer to bind DNA sequences containing that feature. Alternatively, if the structural feature were merely a byproduct of the DNA sequences selected by a base readout mechanism in the major groove, the binding sites preferred by the mutant factor would still contain that feature. Here, we tested this prediction using the anterior Hox protein Scr, which binds DNA with Exd to regulate Scr-specific target genes during *Drosophila* embryogenesis (Ryoo and Mann, 1999). In a co-crystal structure of the Exd-Scr heterodimer bound to an Scr-specific target site, *fkh250* (AGATTAAT), both shape readout and base readout mechanisms were evident (Joshi et al., 2007). In agreement with the SELEX-seq data, the *fkh250* binding site contained two MG width minima, one recognized by Scr residues His-12 and Arg3, and the second recognized by Scr residue Arg5 (Figures 1A and 1B) (Joshi et al., 2007). As these residues did not form hydrogen bonds with bases, the implication is that they use shape readout, and not base readout, as their sole mode of DNA recognition.

To test if Hox proteins directly use shape readout, we characterized the properties of mutant proteins that, based on the Exd-Scr co-crystal structures, are predicted to either lose or gain the ability to read specific MG topographies. When MG-inserting residues of Scr were mutated to alanines, thus impairing its ability to use shape readout, the mutant proteins no longer preferred sequences containing these MG width minima. Conversely, when MG recognizing residues from Scr were transferred to a Hox protein that normally does not select for this structural feature, the proteins selected binding sites with two MG minima in vitro and gained the ability to activate an Scr-specific target gene in vivo. Finally, we show that taking DNA shape features into consideration significantly improved the ability to predict Exd-Hox binding site specificities compared to models that only depend on DNA sequence. Together, these findings demonstrate that transcription factors directly use shape readout for protein-DNA recognition, and in silico prediction of DNA binding specificities will benefit by taking DNA structural features into consideration.

## RESULTS

### Mutants that Interfere with Scr's Ability to Read MG Shape

In an initial set of experiments to tease apart the contributions of shape readout from base readout, we mutated Scr residues His-12, Arg3, and Arg5, which, in a co-crystal structure, only use shape readout as their mode of recognition (Joshi et al., 2007) (Figures 1A and 1B). We generated a series of mutant proteins that change these residues to alanines and, consequently, impair Scr's ability to recognize local MG topographies. We mutated either Arg3 alone ($Scr^{Arg3A}$), His-12 alone ($Scr^{His\text{-}12A}$), both His-12 and Arg3 ($Scr^{His\text{-}12A,\ Arg3A}$), or Arg5 alone ($Scr^{Arg5A}$) and tested the effect of these mutations in complex with Exd on Scr's DNA binding site preferences using SELEX-seq (Figure 1C).

Because the binding site for Exd-Hox complexes is 12 base pairs (Slattery et al., 2011), we generated 12-mer relative affinities for each Scr mutant in complex with Exd using small modifications of our previously described procedure (see Experimental Procedures) (Riley et al., 2014; Slattery et al., 2011), and compared them to the affinities generated by wild-type (WT) Exd-Scr heterodimers. We color-coded the 12-mers based on their core 8-mer (Figure 1D) (Slattery et al., 2011). Compared to Scr WT, all three mutants showed an increased relative preference for the green (TGATTGAT), yellow (TGATGGAT), and purple (TGATCGAT) motifs, and a decrease in the preference for the blue (TGATTAAT) motif (Figures 1D and 1E). Because the blue motif includes the Scr-specific *fkh250* binding site, we directly compared the blue and red (Exd-Hox consensus) motifs by plotting the relative affinities of 12-mer pairs that only differed at the single position that distinguished them from being blue or red (e.g., nnTGATT**A**ATnn with nnTGATT**T**ATnn). Whereas ScrWT showed a preference for blue compared to red motifs over the entire range of affinities, this preference was weakened for $Scr^{Arg3A}$ and abolished for $Scr^{His\text{-}12A,\ Arg3A}$ (Figure 1F).

Although the above results show that Arg3 and His-12 are required for Scr's binding site preferences, they do not address if this is due to their preference for a specific MG shape. To determine if His-12, Arg3, and Arg5 directly enable the selection of sequences with narrow MGs, we computed the average MG width profile for thousands of 16-mer sequences that were preferentially bound by each Scr variant in our SELEX-seq experiments. We employed DNAshape, a high-throughput method for the prediction of the structural features of DNA sequences based on the average conformations of pentamers derived from all-atom Monte Carlo simulations (Zhou et al., 2013). Sequences selected by $Scr^{His\text{-}12A,\ Arg3A}$ had an average MG width at $A_9$ and $Y_{10}$ that was significantly wider compared to those selected by ScrWT, without affecting the selection of the MG width minimum at $A_5Y_6$ (Figures 2, S2, and S3) ($p < 2 \times 10^{-16}$; Mann-Whitney U test). Sequences selected by the single mutant $Scr^{Arg3A}$, but not $Scr^{His\text{-}12A}$, had an intermediate width at $A_9Y_{10}$, suggesting that His-12 and Arg3 synergistically contribute to MG recognition at the Hox half-site (Figures 2 and S3). Conversely, compared to ScrWT, $Scr^{Arg5A}$ selected sequences with a wider

(E) Comparative specificity plots comparing the relative binding affinities of 12-mers selected by Exd-ScrWT (y axis) with each Exd-Scr variant (x axis). 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 most common cores. The black line indicates y = x.

(F) Plots comparing the relative affinities of sequences containing a blue motif (TGATTAAT) (y axis) versus a red motif (TGATTTAT) (x axis) for Exd-ScrWT and Exd-Scr variants. Each point represents the relative affinities of a pair of 12-mers that are identical except for the position that makes it either a blue (nnTGATT**A**ATnn) or a red (nnTGATT**T**ATnn) motif. The black line indicates y = x, and the red line is a linear regression trend line. The slope of the trend line and coefficient of determination $R^2$ of the data are indicated.

See also Figures S1, S2, and S3.

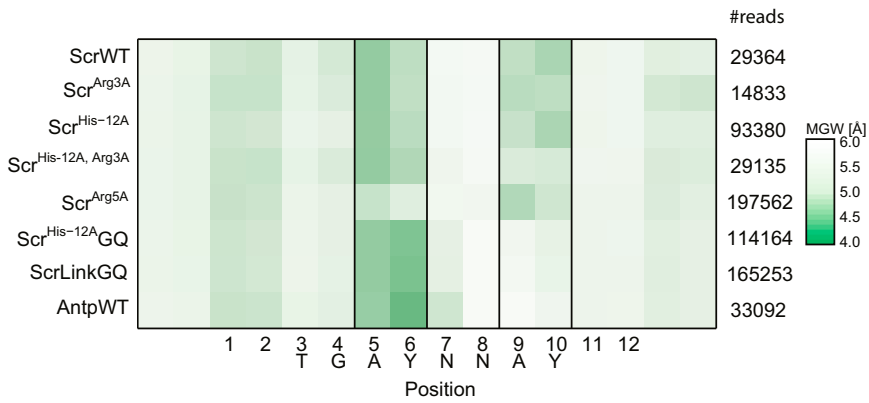


**Figure 2. Loss of MG Width Preferences in the Absence of MG-Recognizing Residues**

Heat map of the average MG width at each position of 16-mers selected by each Exd-Hox heterodimer. Dark green represents narrow MG regions whereas white represents wider MG regions. The number of sequences analyzed for each complex is shown on the right. Black lines demarcate where Arg5 inserts into the MG ($A_5Y_6$) and, for ScrWT, where Arg3 and His-12 insert into the MG ($A_9Y_{10}$).

MG specifically in the Exd half-site ($A_5Y_6$), but these sequences retained the minimum at $A_9Y_{10}$ (Figures 2 and S3). These results provide strong support for the idea that Arg5 directly selects sequences with a MG minimum at $A_5Y_6$, while Arg3/His-12 directly select sequences with a MG width minimum at $A_9Y_{10}$. Selection of these MG width minima occurs independently, even though they are only separated by two base pairs.

Despite its importance in selecting Scr-specific features of MG topography, Arg3 is present in many Hox homeodomains, including Antennapedia (Antp), that do not select a MG minimum at $A_9Y_{10}$ (Figures 3A and S1) (Slattery et al., 2011). This observation prompted the question of why Arg3 in Antp and other posterior Hox proteins does not select for a narrow MG at this position. We speculated that the amino acids flanking Arg3 might play a role in binding site selection by correctly positioning this MG-inserting side chain. Indeed, although both Scr and Antp have Arg3 and Arg5, these residues are part of an N-terminal arm motif that differs between these two Hox proteins ($R_3Q_4R_5T_6$ in Scr and $R_3G_4R_5Q_6$ in Antp) (Figure 3A). To test whether residues flanking these arginines play a role in Scr binding specificity we characterized an additional mutant, $Scr^{His-12A}GQ$ (Figures 1C–1F). In $Scr^{His-12A}GQ$, His-12 is mutated to alanine and the fourth and sixth positions in the Scr homeodomain are changed to that of Antp ($Gln_4$ to $Gly_4$ and $Thr_6$ to $Gln_6$) to mimic Antp's $R_3G_4R_5Q_6$ motif. Strikingly, this mutant failed to select sequences with a minimum at the Hox half site ($A_9Y_{10}$) (Figure 2). An additional mutant, ScrLinkGQ, that, in addition to having Antp's $R_3G_4R_5Q_6$ motif, has Antp's linker (residues in between the YPWM Exd interaction motif and the homeodomain, Figure S2A) in place of Scr's linker, showed very similar behavior to $Scr^{His-12A}GQ$ (Figures 2 and S2). Together, these data suggest that additional residues within and adjacent to the N-terminal arm, which do not make direct contact with the DNA (minor or major groove), play an important role in selecting Hox-specific MG topographies, likely by positioning the MG inserting side chains of Arg3, Arg5, and His-12.

## Mutants that Transfer Scr's Ability to Read MG Shape to Antp

The above experiments demonstrate that MG-inserting side chains in Scr are necessary for Scr's ability to select sequences with local MG width minima. To test whether MG recognizing residues are sufficient to confer Scr's binding preferences to a different Hox protein, we introduced these residues into Antp, which normally prefers sequences with wider MG regions at $A_9Y_{10}$ (Figure 2). We created a series of Antp mutants that contained various combinations of Scr-specific amino acids in two regions, the linker and the N-terminal arm motif $R_3Q_4R_5T_6$ (Figure 3A). Remarkably, the 12-mer relative affinity profiles of these Antp mutants (in complex with Exd) gradually converged toward that of ScrWT upon the introduction of residues used for MG width recognition (Figures 3B–3D). All of the residues tested—Gln4, Thr6, His-12, and the linker—contributed to the convergence of Antp's binding specificity toward that of Scr, with the most Scr-like mutant, AntpLinkQT, being nearly indistinguishable from ScrWT (Figures 3B and 3C). A direct comparison of the relative affinity for the red motif versus the Scr-preferred blue motif also revealed a gradual shift in preference toward the blue motif (Figure 3D). Thus, Scr-specific amino acids from its linker and N-terminal arm are sufficient to confer Scr's binding specificity on another Hox protein.

To determine if these Antp mutants also share Scr's MG shape preferences, we used DNAshape to predict the MG widths of 16-mers selected by these proteins. In general, the average MG width at $A_9Y_{10}$ of the sequences selected by the Antp mutant series became narrower, toward that of Scr, upon the introduction of Scr-specific residues (Figure 4A), where, with the exception of AntpHQ, each successive mutant selected sequences with a statistically significant narrowing of the average MG at these positions (Figure S4A). On average, the differences in MG widths at these positions were larger for high-affinity sequences than for low-affinity sequences (Figure S4B). Taken together, these results suggest that Scr residues Gln4, Thr6, His-12, and linker all contribute to the recognition of DNA shape. Moreover, these residues are sufficient to confer the shape preferences of Scr when inserted into another Hox protein.

As an alternative way to analyze these data, we compared the binding specificities of each Antp variant geometrically by calculating the Euclidean distances between the MG width profiles of sequences selected by each variant with the average MG width of those selected by Exd-AntpWT and Exd-ScrWT, respectively. The resulting density plots showed two occupancy peaks, one representing sequences that are more similar to those selected by AntpWT, and a second representing sequences that are more similar to those selected by ScrWT (Figure 4B). With the exception of AntpHQ, each successive Antp variant showed a gradual shift toward the ScrWT peak, with AntpLinkQT showing

a nearly complete shift. Thus, key Scr-specific residues in the linker and N-terminal arm were sufficient to convert Antp's shape preferences to those of Scr.

### Antp Variants that Mimic Scr's DNA Shape Preferences Activate an Scr-Specific Target In Vivo

The above results demonstrate that shape readout, mediated by a limited number of Hox residues, is an essential component of DNA recognition by Exd-Hox heterodimers in vitro. But how relevant is this readout mechanism in vivo? To answer this question we examined the ability of these Antp variants to activate *fkh250-lacZ*, an Scr-specific reporter gene that contains a binding site (AGATTAAT) with two MG width minima (Joshi et al., 2007). In otherwise wild-type embryos, *fkh250-lacZ* expression was limited to Scr-expressing cells in parasegment 2 (PS2) (Ryoo and Mann, 1999) (Figure 5A). Ectopic expression of ScrWT using the *prd-Gal4* driver activated *fkh250-lacZ* in segments outside PS2 (Figure 5B), and His-12 and Arg3 of Scr are required for this activation (Joshi et al., 2007). In contrast to Scr, ectopic expression of AntpWT did not activate *fkh250-lacZ* (Figure 5C). However, ectopic expression of AntpHQT resulted in modest *fkh250-lacZ* activation (Figure 5D), while ectopic expression of AntpLinkQT, the Antp mutant whose binding specificity most closely resembled Scr in vitro (Figures 3 and 4), resulted in strong activation of *fkh250-lacZ* (Figure 5E). Thus, Antp mutants that prefer to bind sequences with two MG width minima in vitro, the normal topography of an Scr-specific binding site, also have the ability to activate an Scr-specific target gene in vivo.

### DNA Shape Features Improve Accuracy of Binding Specificity Predictions

If shape readout is a direct and independent determinant of Hox-DNA binding specificity, we speculated that shape features of the target DNA could be used to improve quantitative predictions of relative binding affinities. To test this notion, we trained an L2-regularized multiple linear regression (MLR) model (Yang et al., 2014) for each of the mutants and WT Hox proteins. We used 10-fold cross validation in order to train and determine the accuracy of a given model, quantified as the coefficient of determination $R^2$. These MLR-derived $R^2$s are robust as they are highly correlated with $R^2$s derived using an alternative machine learning approach, support vector regression ($\varepsilon$-SVR) with a linear kernel (Figure S5; see Experimental Procedures for details) (Gordân et al., 2013; Zhou et al., 2015).

Using MLR, addition of MG width to a model based only on nucleotide sequence resulted in a modest improvement in $R^2$ of on average 12% (Figures 6A and S6A). Like MG width, adding three other shape features one at a time, Roll, propeller twist (ProT), and helix twist (HelT), also led to a modest improvement in accuracy (Figure S6A). Inclusion of all four DNA shape features in combination further increased prediction accuracy (Figures 6B and S6A). The improvement in binding affinity prediction accuracy, on average 26% when incorporating all four shape features, yielded the largest effect with high significance ($p = 6 \times 10^{-5}$; Mann-Whitney U test). The addition of any combination of three shape features led to an intermediate increase in prediction accuracy, in some cases similar to that after addition of all four shape features (Figure S6A). These results suggest that all four features contribute to Exd-Hox-DNA target selection in a non-additive manner, consistent with the interdependency of these features (Olson et al., 1998). Thus, including DNA shape features in addition to MG width improves binding site predictions over models based only on nucleotide sequence.

For comparison, we also assessed the benefit of adding shape features for the prediction of Hox monomer specificities. Interestingly, in this case the improvement in $R^2$ was, on average, only 6.4%, suggesting a larger role for DNA shape in conferring heterodimer specificity than monomer specificity (Figure S6B).

### DNA Shape Contributes to Binding Specificities in a Position-Specific Manner

Next, we hypothesized that if shape features contribute to an improvement in binding specificity prediction, then it might be possible to localize this effect within the binding site. We trained models using the sequence of the entire binding site augmented by all four shape features at individual positions one at a time, resulting in a set of models that tested the contribution of shape at each position of the binding site. We compared these models to a sequence-only model and calculated a $\Delta R^2$. This analysis highlighted the importance of DNA shape for predicting Exd-Hox binding specificities in the core, but not the flanks, of the binding site (Figure 6C).

To analyze the role of DNA shape in a complementary manner, we trained shape-only models using the four shape features at all nucleotide positions, leaving out this information one position at a time, resulting in a set of models that assessed the relative importance of DNA shape at each position of the binding site. $\Delta R^2$s were calculated relative to a model that included the four shape features at all positions. In this analysis, prediction accuracy was expected to decrease most when shape features were removed from the model at positions that were important for shape readout. Interestingly, we detected the greatest effect at the $A_9$ position of the Hox half-site, followed by slightly weaker effects at the adjacent $Y_{10}$ position and the $G_4$ position of the Exd half-site (Figure 6D). Eliminating shape features from the remaining positions had a smaller impact on the ability to predict binding specificities.

Within each SELEX-seq data set, the sequences were most variable at the $N_8$ position, raising the possibility that the success of these models might be driven in large part by this position. To test this idea and better assess the role of DNA shape throughout the binding site we trained additional models in which we removed sequence information at the $N_8$ position ("sequence-$N_8$ model"). Leaving out sequence information at the $N_8$ position did not significantly affect the accuracy of a sequence+shape model, suggesting that sequence information at $N_8$ is not essential for its performance (Figure S7A). When MG width information was added to the sequence-$N_8$ model, the ability to predict binding specificities was greatly enhanced compared to the same model without MG width information (Figures S7B and S7C). These results argue that MG width information is more important than sequence at positions with a degenerate sequence signal, such as at $N_8$, where direct readout is not playing a role. The removal of the confounding sequence information at this position uncovered MG width as an independent specificity determinant.

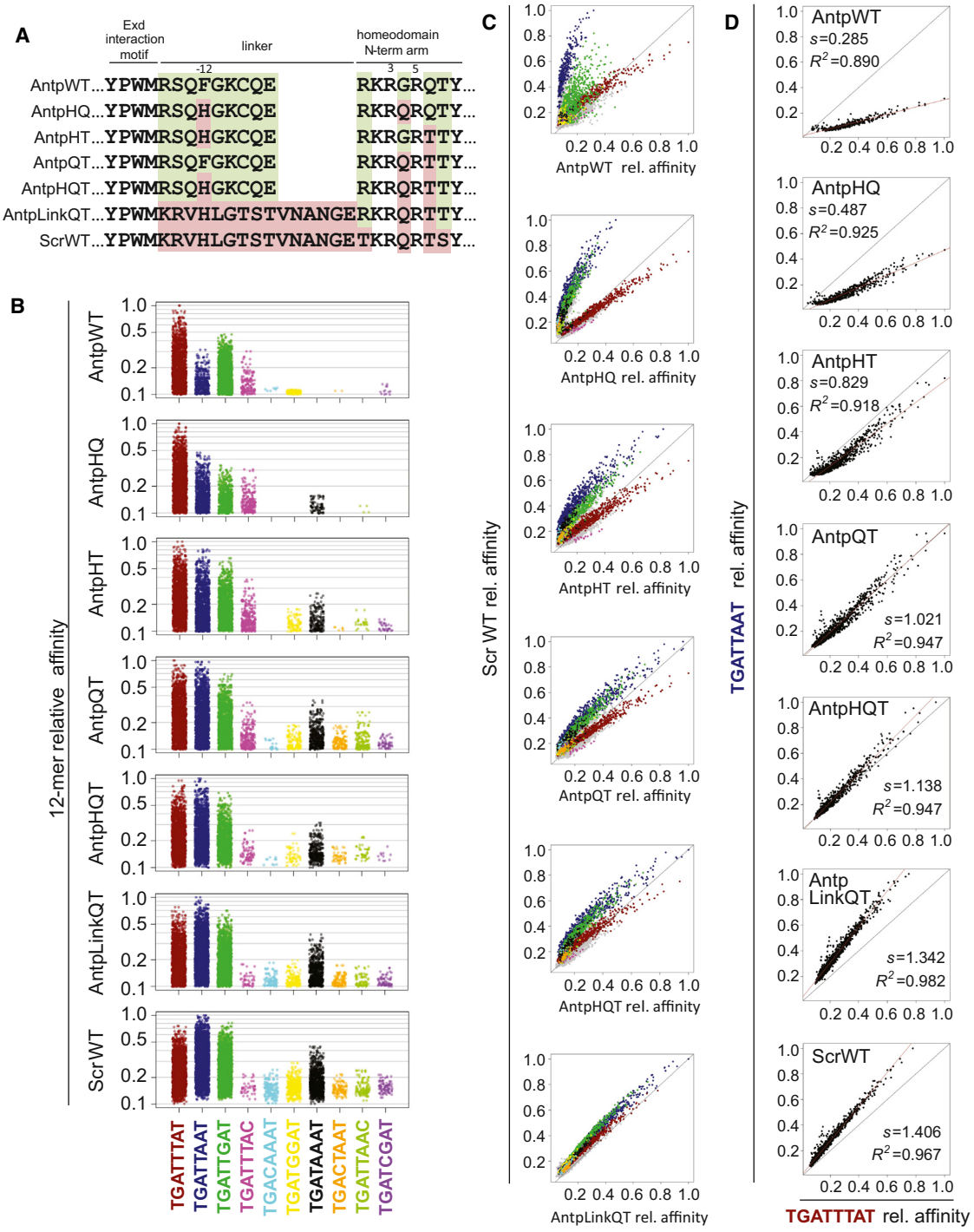


**Figure 3. Introducing Scr's MG Width-Recognizing Residues into Antp Converts Its Binding Specificity to that of Scr**

(A) Amino acid sequences (from the Exd interaction motif, YPWM, through the N-terminal arm of the homeodomain) of Antp variants. Green highlights residues specific to AntpWT, and red highlights residues specific to ScrWT. Non-highlighted residues are common between the two Hox proteins. Numbering is relative to the first residue of Scr's homeodomain. The rest of the protein is wild-type in all cases.

(B) 12-mer relative affinities of binding sites selected by each Antp variant in complex with Exd are color-coded according to the ten most commonly observed Exd-Hox motifs. AntpWT and ScrWT are included to show the progression of the binding preferences from AntpWT toward ScrWT.

(C) Comparative specificity plots of the relative affinity of sequences selected by Exd-ScrWT (y axis) and each Exd-Antp mutant (x axis). 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 most common cores. The black line indicates y = x.

*(legend continued on next page)*

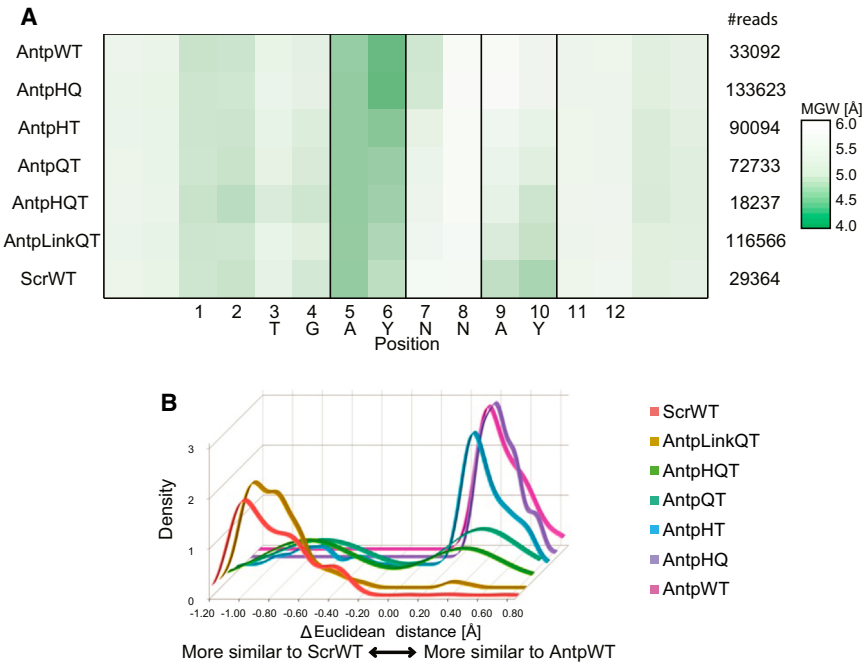


**Figure 4. Shape Readout Properties of Antp Variants with Scr-Specific Residues**

(A) Heat map of the average MG width at each position of all statistically significant 16-mers selected by each Exd-Hox complex. Dark green represents narrow MG regions whereas white represents wider MG regions. The number of sequences analyzed for each protein is shown on the right. Black lines demarcate where Arg5 inserts into the MG ($A_5Y_6$) and, for Scr, where Arg3 and His-12 insert into the MG ($A_9Y_{10}$).

(B) Histogram representing the distribution of MG width similarities for each of the sequences selected by each Antp variant in comparison to those selected by ScrWT and AntpWT. The y axis represents the density of 16-mers at different Δ(Euclidean distance) scores (x axis). Sequences more similar to those selected by ScrWT receive a negative score, and sequences more similar to those selected by AntpWT receive a positive score. See also Figure S4.

When all four shape features were added to the sequence-$N_8$ model at single positions one at a time the contribution of DNA shape within the core motif was very apparent (Figure 7A), and significantly stronger than when the starting model included sequence information at the $N_8$ position (compare with Figure 6C). If instead of all four DNA shape parameters only MG width was added position by position to the sequence-$N_8$ model, the average improvement in $R^2$, while smaller, was most apparent at or adjacent to $Y_6N_7$ and $A_9Y_{10}$ (Figure 7B). Thus, although DNA shape is generally important within the entire core of the binding site, the contribution of MG width is strongest at the two AY regions, precisely where local minima in MG width were observed in the Exd-Hox X-ray structures (Joshi et al., 2007) and SELEX-seq data (Figures 2 and 4).

Taken together, quantitative predictions based on regression models indicated that shape features become important where sequence information is not well defined, more likely at positions that are not involved in base readout. In these cases, shape features contain more information than sequence alone, and removing the signal from sequence enables the quantitative modeling of the role of shape features on binding specificity.

## DNA Shape Features Discriminate Anterior from Posterior Hox Binding Specificities

To understand to what extent shape features can help distinguish Exd-ScrWT from Exd-AntpWT binding specificities, we assigned a value of +1 to the top 50% of sequences selected by Exd-ScrWT and −1 to the top 50% of sequences selected by Exd-AntpWT (see Experimental Procedures for details). We then used sequence- and shape-based models to evaluate the discriminative power of the selected features. Using L2-regularized MLR and 10-fold cross validation, we calculated the area under the receiver-operating characteristic curve (AUC) as a criterion for a model to discriminate ScrWT-like from AntpWT-like binding specificities. We found that MG width alone, without using sequence or additional shape features, discriminates between the binding specificities of both Exd-Hox complexes with high accuracy (Figure S7D). Thus, MG width does not merely refine binding specificity but is a powerful descriptor on its own, at least for discriminating between these two Exd-Hox complexes. Classification models using other shape parameters performed similarly well (Figure S7D), indicating that a classification between two states is less sensitive than quantitative prediction of binding strength using regression models. Further, these results suggest that the qualitative differences that are apparent in the MG width heat maps (Figures 2 and 4A) reflect a quantitative difference in anterior and posterior Hox specificities.

Next, we asked which positions in the binding site had the highest impact on this classification. To answer this question, we calculated the Pearson correlation between the class labels +1 and −1 for Exd-ScrWT and Exd-AntpWT, respectively, and MG width at each position (see Experimental Procedures for details). Several positions showed strong, either positive or negative, correlations that enabled the classification into ScrWT-like and AntpWT-like binding specificities (Figure 7C). Two regions showing a negative Pearson correlation aligned with the two MG width minima observed in the Exd-Scr co-crystal structure, and a region of positive Pearson correlation marked

(D) Plots comparing the relative affinities of sequences containing a blue motif (TGATTAAT) (y axis) versus a red motif (TGATTTAT) (x axis) for ScrWT, AntpWT and Antp variants. Each point represents the relative affinities of a pair of 12-mers that are identical except for the position that makes it either a blue (TGATT**A**AT) or a red (TGATT**T**AT) motif. The black line indicates y = x, and the red line is a linear regression trend line. The slope of the trend line and coefficient of determination $R^2$ of the data are indicated.

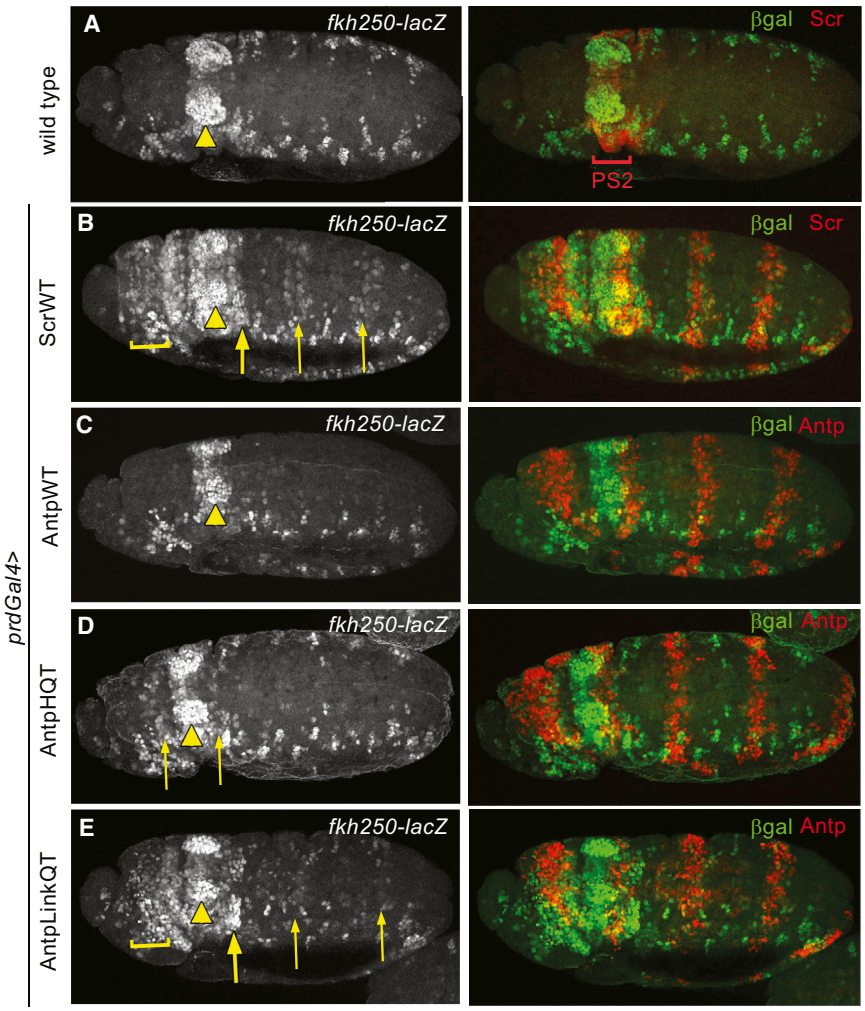


**Figure 5. Scr's MG Width Readout Residues Confer the Ability to Activate an Scr-Specific Target In Vivo when Incorporated into Antp**

(A) In wild-type embryos *fkh250-lacZ* is activated only in parasegment 2 (PS2), where endogenous Scr is expressed (arrowhead). In this and all panels, anterior is to the left.

(B) Ectopic expression of ScrWT using *prd-Gal4* (visualized with red stripes of ectopic expression in the panel on the right) activates *fkh250-lacZ* anterior and posterior to PS2. Activation is strongest anterior to PS2 (bracket) and immediately posterior to PS2 (thick arrow), with weaker activation in abdominal segments (thin arrows).

(C) Ectopic expression of wild-type Antp does not activate *fkh250-lacZ*.

(D) Ectopic expression of AntpHQT leads to weak ectopic *fkh250-lacZ* expression anterior and posterior to PS2 (thin arrows).

(E) Ectopic expression of AntpLinkQT leads to activation both anterior and posterior to PS2. Activation is strongest anterior to PS2 (bracket) and immediately posterior to PS2 (thick arrow), with weaker activation in abdominal segments (thin arrows).

the region between these minima. This observation confirms that the core region is important for the differences in binding specificity between paralogous Hox factors. Interestingly, not only is the AY region of the Hox half-site important, but the shape of the entire core, presumably due to the influence of all core positions on the shape of this region.

Finally, we used classification models to predict whether the DNA shape mutants defined in Figures 1, 3, and S2 tend to show ScrWT-like or AntpWT-like binding specificities. Here, a sequence was classified as ScrWT-like if the class label was predicted to be >0, and as AntpWT-like if the class label was predicted to be <0. This classification indicated a gradual change in the fraction of sequences selected by any of the mutants assigned as ScrWT- versus AntpWT-preferred sequences (Figure 7D). These data quantitatively confirm the qualitative observations shown above (Figures 2 and 4A) that MG width topography is an important binding specificity signal for Hox proteins.

## DISCUSSION

Despite significant effort in the field, it is still not possible to accurately decipher the regulatory information that is encoded in the DNA sequences of eukaryotic genomes (Slattery et al., 2014). In the work described here, we used a combination of in vitro, in vivo, and computational approaches to show that intrinsic DNA structural characteristics—collectively referred to as DNA shape—are being directly read by DNA binding proteins when they recognize their binding sites. Thus, analogous to mechanisms in which DNA base pairs are directly read by proteins via hydrogen bonds, the recognition of DNA shape independently contributes to both binding affinity and specificity. Using this information, we show that including DNA shape features significantly enhances the ability to predict DNA binding specificities and thus will greatly improve models for accurately predicting transcription factor binding in eukaryotic genomes.

### Separable Contributions of DNA Shape and Sequence to Protein-DNA Recognition

Although several previous reports suggested the importance of DNA shape in protein-DNA recognition, all prior work was unable to definitively discriminate between the roles of DNA shape and sequence. Although DNA shape features, such as MG width, were previously found to contribute to binding specificity (Dror et al., 2014; Gordân et al., 2013; Lazarovici et al., 2013; Yang et al., 2014), here the roles of DNA sequence and shape have been separated and analyzed in an unbiased manner. To achieve this, we mutated Scr amino acid side chains that do not make direct base contacts in the major groove, but instead either insert into the MG (His-12, Arg3, Arg5) or indirectly influence these interactions (Gln4, Thr6, linker). The combination of SELEX-seq

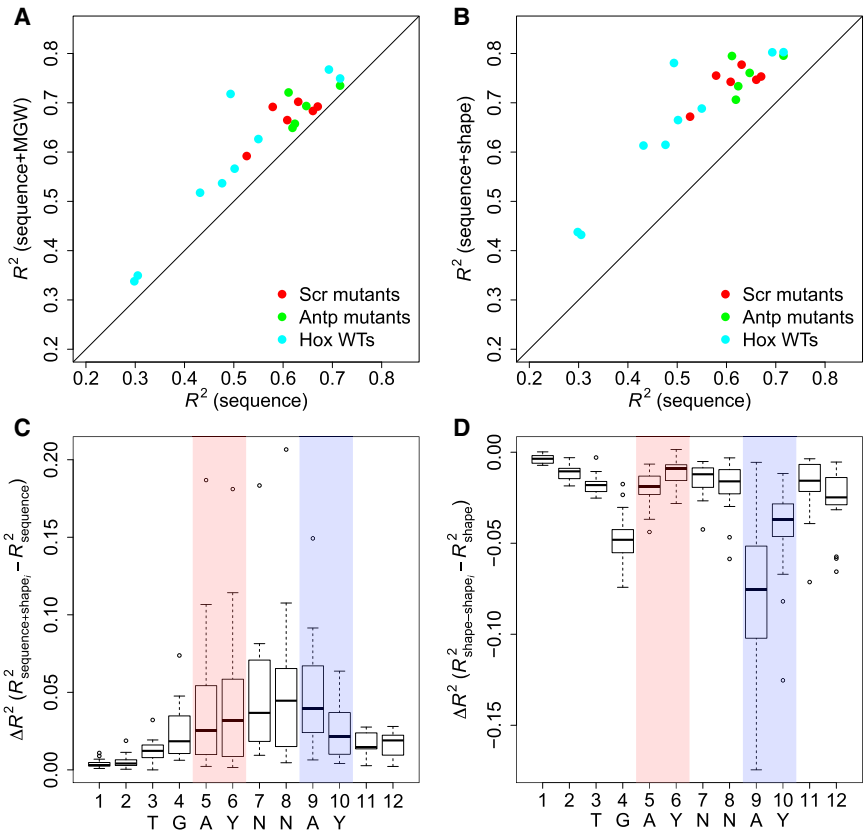


**Figure 6. DNA Shape Features Improve Quantitative Predictions of DNA Binding Specificities of Exd-Hox Heterodimers**

(A) Scatter plot representing the coefficient of determination $R^2$ obtained using a sequence-only model (x axis) compared to a model using sequence and MG width (y axis). Each point represents a different Exd-Hox heterodimer and is color-coded as indicated.

(B) Scatter plot representing the coefficient of determination $R^2$ obtained using a sequence-only model (x axis) compared to a model using sequence and four DNA shape features (MG width, Roll, ProT and HelT) (y axis). Quantitative measures for the improvement of the prediction accuracy of the logarithm of relative binding affinities using shape-augmented models are provided in Figure S6.

(C) Box plots illustrating the contribution from DNA shape features to model accuracy when shape features were added to a sequence model at each position individually. The effect on the coefficient of determination $\Delta R^2$ is shown for adding four shape features (MG width, Roll, ProT and HelT) position-by-position to the sequence model. The centerline of the box plots represents the median, the edge of the box the first and third quartile, and the whiskers indicate minimum/maximum values within 1.5 times the interquartile from the box.

(D) Box plots illustrating the contribution from DNA shape features to model accuracy when sequence features were removed. The effect on the coefficient of determination $\Delta R^2$ is shown for leaving out four shape features (MG width, Roll, ProT, and HelT) position-by-position from a shape-only model that does not contain any sequence information. The box plots are defined in (C).

See also Figures S5 and S6.

with high-throughput DNA shape analysis allowed us to show the effect of these mutations on the selection of DNA binding sites with distinct shape characteristics. Further, not only were these amino acid side chains necessary for conferring the DNA binding preferences of these proteins, they were sufficient to confer this specificity, both in vitro and in vivo, when grafted into a different Hox protein, Antp. These experiments effectively tease apart the contributions of shape readout from base readout. We speculate that the readout of DNA shape may be a general mechanism that transcription factors use to recognize their binding sites. Moreover, for transcription factors that are members of large paralogous families, such as the Hox proteins, DNA shape may be essential for distinguishing between binding sites that are difficult to discriminate based on base readout alone.

## Statistical Machine Learning Reveals DNA Structure-Based Binding Specificity Signals

To complement and extend the in vitro and in vivo studies, we used statistical machine learning, in this case multiple linear regression (MLR), to computationally analyze the contributions of DNA sequence and shape. Using this approach we were able to (1) quantify the overall contribution of shape features to binding specificity and (2) compute the relative contributions of DNA shape and sequence at individual positions within the binding site. Extensive experimental work, involving structure determination and mutagenesis, represents the current standard approach for uncovering DNA readout mechanisms of transcription factors. The quantitative modeling introduced here suggests an alternate route for deriving such mechanistic information from high-throughput sequencing data. These methods will therefore likely be valuable when used to predict the DNA binding specificities of other transcription factors and when analyzing their interactions with genomes.

To identify positions in the binding site where shape features contribute substantially to binding specificity, we used a form of feature selection in which we compared models with different feature sets by computing a $\Delta R^2$ relative to a reference model. We found that the shape features in the core of the Exd-Hox heterodimer binding site were important for paralogous binding specificity. This observation is distinct from previous observations for another family of transcription factors, basic helix-loop-helix (bHLH) factors, where shape features in regions flanking the core binding site play an important role in discriminating binding specificities of related family members in yeast (Gordân et al., 2013) and human (Yang et al., 2014). Further, our feature selection approach indicates that shape features at the AY region of the Hox half-site were the most critical for determining binding specificity. This finding agrees with qualitative

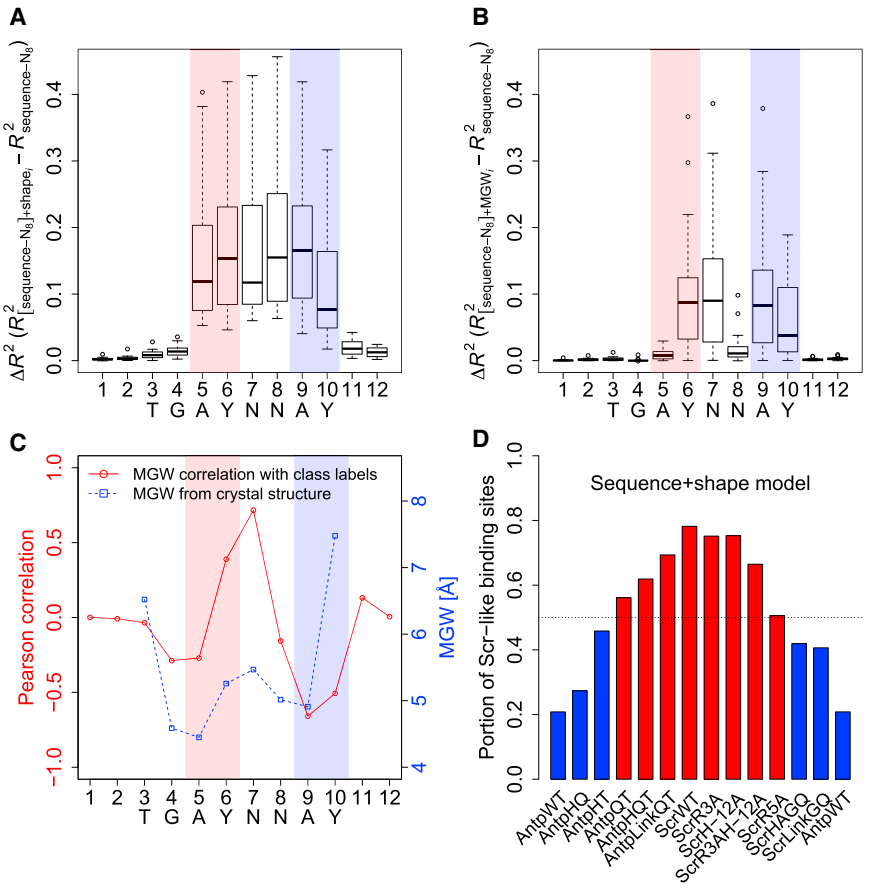


**Figure 7. Models that Deconvolve DNA Sequence and Shape**

(A) Removing sequence features at the $N_8$ position where sequence is least constrained across the selected sequences from the sequence+shape model further emphasizes the contribution of adding DNA shape to model accuracy. Whereas removing sequence information at this position has essentially no effect on model accuracy (Figure S7A), adding MG width to the sequence-$N_8$ model has a large effect on prediction accuracy (Figure S7B). Based on this finding, the effect on the coefficient of determination $\Delta R^2$ is shown in box plots for adding four shape features (MG width, Roll, ProT, and HelT) position-by-position to the sequence-$N_8$ model. The centerline of the box plots represents the median, the edge of the box the first and third quartile, and the whiskers indicate minimum/maximum values within 1.5 times the interquartile from the box.

(B) Box plots illustrating the effect on the coefficient of determination $\Delta R^2$ for adding MG width information position-by-position to the sequence-$N_8$ model emphasize the role of the AY and immediately adjacent positions. The box plots are defined in (A).

(C) Pearson correlations (red) between MG width (MGW) and binding site labels (+1 for ScrWT-like versus −1 for AntpWT-like) track with the MGW pattern (blue) observed in the co-crystal structure (Joshi et al., 2007), emphasizing the important role of MGW in the core region of Exd-Hox binding site.

(D) A sequence+shape classification model captures the gradual change of binding specificities introduced by mutations of the N-terminal arm and linker sequences with some Exd-Hox mutant heterodimer specificities classified as Scr-like (red) and others as Antp-like (blue).

See also Figure S7.

observations in a previous study (Slattery et al., 2011) and in this work (Figures 2 and 4) that shape selections varied most substantially at this position for both wild-type and mutant Hox proteins. While this was previously a qualitative observation, the current study shows the effect quantitatively. The machine learning and feature selection methods reveal that this information will likely provide a powerful approach when analyzing data from high-throughput binding assays for other transcription factors. In particular, it is noteworthy that we were able to derive structural mechanisms used by Hox transcription factors based only on sequence data alone, without solving a 3D structure.

## Broader Implications for Recognition of Genomic Target Sites by Transcription Factors

Based on our findings, we propose that as more high-throughput DNA binding data become available (Hume et al., 2015; Jolma et al., 2013; Zhu et al., 2011), DNA shape parameters should be taken into consideration when analyzing and subsequently scanning genomes for DNA binding site preferences. Further, although different families of transcription factors may use DNA shape in various ways, this information may be used to inform binding site prediction algorithms. As shown here quantitatively, Exd-Hox heterodimers use distinct structural features in the DNA, such as local regions of narrow MG, to achieve DNA binding specificity. Because MG width minima are distinct structural motifs, we were able to separate their contributions to DNA recognition both biochemically, by mutating amino acids that recognize these motifs, and computationally, by training models that include or exclude specific subsets of DNA features. For other protein families, the contribution of DNA structure may not be as readily separable as it is for Exd-Hox binding. For example, although previous work demonstrated a role for DNA shape in conferring the binding specificity of bHLH proteins, this effect was mediated by sequences flanking the core binding site (E-box), where no known protein-DNA interactions (base or shape readout) occur (Gordân et al., 2013). In this case, the role of DNA shape may be biochemically inseparable from base readout because it is unlikely that a distinct structural motif is formed by the flanking sequences.

Our results have implications for the design of binding site search and de-novo motif discovery methods, which currently most typically rely only on DNA base features (Weirauch et al., 2013). There are some examples where large sets of overlapping DNA structural features, which are highly interdependent from each other and inseparable from sequence, have been integrated in motif search algorithms (Hooghe

et al., 2012; Maienschein-Cline et al., 2012; Meysman et al., 2011). The results described here, however, suggest that for some transcription factor families, distinct structural motifs, which can be defined independently from sequence, such as MG topography, can be directly integrated in genome analysis tools as quantifiable search parameters. The ability to independently define and quantify the role of distinct structural motifs will likely yield more powerful algorithms that may help identify low affinity, high specificity Hox binding sites that are unrecognizable with standard approaches (Crocker et al., 2015). Further, machine learning approaches may also contribute to more accurate models of cooperative transcription factor binding, for example in the interferon-β enhanceosome (Chang et al., 2013), or in vivo, where DNA shape has been identified as a predictive feature for transcription factor binding (Barozzi et al., 2014). We further propose that the computational approaches described here will also be valuable for deconvolving and discovering the roles of DNA shape and sequence even for transcription factors such as the bHLH factors where DNA shape cannot be as readily separated biochemically from DNA sequence. The ability to quantitatively assess the distinct roles of DNA sequence and shape will therefore advance our ability to identify bona fide genomic binding sites and the ability to interpret eukaryotic genomes.

## EXPERIMENTAL PROCEDURES

### Oligonucleotides

All oligonucleotides used in this study are listed in Table S1.

### Protein Purification

Scr and Antp mutants were cloned using the QuickChange Site-Directed Mutagenesis Kit (Agilent) using his-tagged ScrWT (Joshi et al., 2007) and his-tagged AntpWT (Jaffe et al., 1997; Noro et al., 2006) as templates. His-tagged proteins were expressed in BL21 cells and purified using Cobalt chromatography. For the SELEX-seq experiments, "Exd" refers to Exd co-purified with the HM domain of Homothorax (Hth) (Noro et al., 2006).

### In Vivo Analysis

All transgenic UAS lines were generated using the ϕ-C31 integration system into the attP2 insertion site. UAS lines were crossed to flies containing *fkh250-lacZ* on the second chromosome and *prd-Gal4* on the third chromosome. Embryos were collected at 25°C and stained using rabbit anti-β-galactosidase (Cappell) and either mouse anti-Scr (gift from D. Andrews) or mouse anti-Antp (8C11; DSHB).

### SELEX-Seq

All SELEX experiments were carried out as described (Riley et al., 2014; Slattery et al., 2011). In total, five 16-mer libraries were used for multiplexing (Table S1). Sequencing was performed by Illumina HiSeq 2000/2500. The number of sequences analyzed for each protein is listed in Tables S2 and S3.

### Inferring Relative Binding Affinities

Fifth order Markov models were constructed using Round 0 (R0) sequences to predict the number of 12-, 14-, and 16-mer sequences in each initial library as described (Riley et al., 2014; Slattery et al., 2011). R3 data were used for all Hox variants in order to optimize counts and minimize sampling error. 12-, 14-, and 16-mer relative binding affinities were generated by taking the cubic root of the enrichment ratio (counts in R3 divided by expected counts as predicted using Markov model derived from R0 data).

### 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. Four DNA structural features were derived for these sequences from a high-throughput DNA shape prediction method (Zhou et al., 2013). Euclidean distance was used to compare MG width profiles of sequences selected by Hox mutants to the average MG width at all positions of sequences selected by the Hox WTs. See Extended Experimental Procedures for details.

### Regression Models for Predicting Binding Specificities Quantitatively

To predict the relative binding affinity for sequences bound by the Hox monomers and Exd-Hox heterodimers, we trained L2-regularized multiple linear regression (MLR) models (Yang et al., 2014). A 10-fold cross-validation was performed with an embedded 10-fold cross-validation on the training set to determine the optimal λ parameter. We trained models that (1) encoded the nucleotide sequence of each of the bound sequences as binary features (sequence models), (2) encoded different combinations of the DNA shape features MG width, ProT, Roll, and HelT (shape models), and (3) combined nucleotide sequence and DNA shape features at the corresponding position (sequence+shape 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. $\Delta R^2$s were defined as described in the text. See Extended Experimental Procedures for details and access to the source code for DNA shape prediction and feature mapping.

### Classification Models for Distinguishing Binding Specificities

To classify Hox binding specificities, we aligned 14-mers selected by Exd-ScrWT (assigned the label +1) or Exd-AntpWT (assigned the label –1) according to the presence of a single core motif TGAYNNAY. We trained classification models using L2-regularized MLR and used the resulting models to classify the top 50% aligned binding sites preferred by the mutants. 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. See Extended Experimental Procedures for details.

## ACCESSION NUMBERS

The Gene Expression Omnibus (GEO) accession number for the SELEX-seq data sets reported in this paper is GSE65073.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.02.008.

## AUTHOR CONTRIBUTIONS

The experiments were conceived by N.A. and R.S.M. The computational approaches were conceived by I.D., L.Y., and R.R. N.A. executed the SELEX-seq experiments with early contributions from M.S. I.D., L.Y., and T.Z. executed the computational analyses. L.Y. designed and implemented the machine learning and feature selection approaches. N.A. and H.J.B. analyzed the SELEX data. N.A., R.S.M., and R.R. wrote the paper.

## ACKNOWLEDGMENTS

We thank Barry Honig, David Stern, and members of the R.R., H.J.B., and R.S.M. laboratories for feedback and comments on this project, and Vince FitzPatrick, Gabriella Martini, Todd Riley, and Roumen Voutev for technical assistance. This work was supported by the NIH (grants R01GM058575 to

R.S.M., F32GM099160 to N.A., R01GM106056 and U01GM103804 to R.R., and R01HG003008 to H.J.B. and R.R.), the USC-Technion Visiting Fellows Program, and an Alfred P. Sloan Research Fellowship (to R.R.).

Received: August 5, 2014
Revised: December 8, 2014
Accepted: January 26, 2015
Published: April 2, 2015

## REFERENCES

Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S., and Natoli, G. (2014). Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. Mol. Cell *54*, 844–857.

Chang, Y.P., Xu, M., Machado, A.C., Yu, X.J., Rohs, R., and Chen, X.S. (2013). Mechanism of origin DNA recognition and assembly of an initiator-helicase complex by SV40 large tumor antigen. Cell Rep. *3*, 1117–1127.

Crocker, J., Abe, N., Rinaldi, L., McGregor, A.P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., et al. (2015). Low affinity binding site clusters confer Hox specificity and regulatory robustness. Cell *160*, 191–203.

Dror, I., Zhou, T., Mandel-Gutfreund, Y., and Rohs, R. (2014). Covariation between homeodomain transcription factors and the shape of their DNA binding sites. Nucleic Acids Res *42*, 430–441.

Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M.L. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep. *3*, 1093–1104.

Hooghe, B., Broos, S., van Roy, F., and De Bleser, P. (2012). A flexible integrative approach based on random forest improves prediction of transcription factor binding sites. Nucleic Acids Res. *40*, e106.

Hume, M.A., Barrera, L.A., Gisselbrecht, S.S., and Bulyk, M.L. (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res *43* (Database issue), D117–D122.

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.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

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.

Kitayner, M., Rozenberg, H., Rohs, R., Suad, O., Rabinovich, D., Honig, B., and Shakked, Z. (2010). Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. Nat. Struct. Mol. Biol. *17*, 423–429.

Lazarovici, A., Zhou, T., Shafer, A., Dantas Machado, A.C., Riley, T.R., Sandstrom, R., Sabo, P.J., Lu, Y., Rohs, R., Stamatoyannopoulos, J.A., and Bussemaker, H.J. (2013). Probing DNA shape and methylation state on a genomic scale with DNase I. Proc. Natl. Acad. Sci. USA *110*, 6376–6381.

Maienschein-Cline, M., Dinner, A.R., Hlavacek, W.S., and Mu, F. (2012). Improved predictions of transcription factor binding sites using physicochemical features of DNA. Nucleic Acids Res. *40*, e175.

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.

Meijsing, S.H., Pufall, M.A., So, A.Y., Bates, D.L., Chen, L., and Yamamoto, K.R. (2009). DNA binding site sequence directs glucocorticoid receptor structure and activity. Science *324*, 407–410.

Meysman, P., Dang, T.H., Laukens, K., De Smet, R., Wu, Y., Marchal, K., and Engelen, K. (2011). Use of structural DNA properties for the prediction of transcription-factor binding sites in Escherichia coli. Nucleic Acids Res. *39*, e6.

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.

Olson, W.K., Gorin, A.A., Lu, X.J., Hock, L.M., and Zhurkin, V.B. (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc. Natl. Acad. Sci. USA *95*, 11163–11168.

Riley, T.R., Slattery, M., Abe, N., Rastogi, C., Liu, D., Mann, R.S., and Bussemaker, H.J. (2014). SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. Methods Mol. Biol. *1196*, 255–278.

Rohs, R., West, S.M., Liu, P., and Honig, B. (2009a). Nuance in the double-helix and its role in protein-DNA recognition. Curr. Opin. Struct. Biol. *19*, 171–177.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009b). 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.

Seeman, N.C., Rosenberg, J.M., and Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. Proc. Natl. Acad. Sci. USA *73*, 804–808.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., et al. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell *147*, 1270–1282.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends Biochem. Sci. *39*, 381–399.

Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., et al.; DREAM5 Consortium (2013). Evaluation of methods for modeling transcription factor sequence specificity. Nat. Biotechnol. *31*, 126–134.

Yang, L., Zhou, T., Dror, I., Mathelier, A., Wasserman, W.W., Gordân, R., and Rohs, R. (2014). TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res. *42*, D148–D155.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. *41*, W56–W62.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordan, R., and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. Proc. Natl. Acad. Sci. USA. Published online March 9, 2015. http://dx.doi.org/10.1073/pnas.1422023112.

Zhu, L.J., Christensen, R.G., Kazemian, M., Hull, C.J., Enuameh, M.S., Basciotta, M.D., Brasefield, J.A., Zhu, C., Asriyan, Y., Lapointe, D.S., et al. (2011). FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. Nucleic Acids Res. *39*, D111–D117.