Cell, Volume *131*

Supplemental Data

Functional Specificity of a Hox

Protein Mediated by the Recognition

# of Minor Groove Structure

Rohit Joshi, Jonathan M. Passner, Remo Rohs, Rinku Jain, Alona Sosinsky, Michael A. Crickmore, Vinitha Jacob, Aneel K. Aggarwal, Barry Honig, and Richard S. Mann

Supplemental Experimental Procedures

Structure determination

The *fkh250* structure was solved first, by molecular replacement using an initial model derived from the Ubx-Exd-DNA complex (using the core 10 bp of DNA) (Passner et al., 1999). The remaining DNA was fit to the electron density, and the structure refined at 2.6 Å resolution (R_{factor} = 23.3%; R_{free} = 30.3%) (Supp. Table 1). The refined structure contains Scr residues 298-384 (residues 314-325 are disordered), Exd residues 242-300, DNA nucleotides 1-40, 112 water molecules. Residues K298, K309, L313, H383, and K384 were modeled as alanines due to the absence of side chain densities. The *fkh250*^{con*} structure was determined by molecular replacement using the solved *fkh250* structure as a search model. The complex, refined at 2.6 Å resolution (R_{factor} = 25.5%; R_{free} = 29.9%), contains Scr residues 298-384 (residues 312-326 are disordered), Exd residues 242-300, DNA nucleotides 1-40, 103 water molecules.

Computational Procedures

All-atom Monte Carlo (MC) simulations were used to predict intrinsic DNA structures (Rohs et al., 2005b). DNA conformation was sampled based on local nucleotide moves combined with an analytic chain closure (Sklenar et al., 2006), an implicit solvent description, explicit sodium counter ions (Rohs et al., 2005b), and the Amber force field (Cornell et al., 1996). Five independent MC simulations were carried out for the *fkh250* and *fkh250*^{con*} sequences started from standard DNA conformations with identical base pair step geometry. The individual MC runs differed in (i) their starting conformations between canonical A- and B-DNA, (ii) their length between 15 and 19 mers, and (iii) the initial sequence of attempted MC moves. Although initial conformations differed by a maximum of 7.5 Å RMSD, the predicted average structures differ just by <1.2 Å RMSD for *fkh250* and <1.0 Å RMSD for *fkh250*^{con*} at a maximum based on 12,400 atoms. The mutant structure predictions are based on a single MC run starting from canonical B-DNA for each of the sequence variants of the 15-mer CTA$\underline{A}$GATT$\underline{A}$AT$\underline{C}$GGC at either one or two of the underlined positions (single and double mutants of *fkh250*). All MC simulations were carried out over 2,000 MC kcycles with the initial 500 kcycles considered as equilibration period (Rohs et al., 2005a).

Electrostatic potentials at geometric midpoints between the O4' atoms of nucleotide *i+1* on the 5' strand and nucleotide *i-1* on the 3' strand were calculated using the DelPhi program (Rocchia et al., 2002). Partial charges and radii were obtained from Amber (Cornell et al., 1996). The calculations were carried out at physiologic ionic strength (*I*=0.145 M). Regions inside the macromolecules were assigned a dielectric constant of ε=2 while the solvent was assigned a dielectric constant of ε=80. A lattice size of 165^3 was used in combination with five focusing steps from 10 to 90 % solute space filling. The final grid spacing was <0.5 Å. Different angular orientations of the DNA on the grid resulted in electrostatic potential differences of <1%.

Protein-DNA binding assays

At the concentrations used, neither Scr nor $\mathrm{Hth}^{\mathrm{HM}}$-Exd bound to *fkh250*, but together they bound cooperatively. The amounts of bound and unbound DNAs were quantified by a phosphorimager. Kds were measured by at least two independent experiments, each containing at least 10 protein concentration data points fit to non-linear hyperbolic curves using ORIGIN software. For the competition experiments, unlabeled *fkh250* was added to the binding reaction at time 0, at the same time labeled *fkh250*^{con} was added.

Table S1: Crystallographic Parameters

Data Collection Statistics[a]**:**

	fkh250	*fkh250*^{con*}
Resolution range (Å)	50 - 2.6	50 - 2.6
Total no. of Reflections	342,322	311,146
No. of unique Reflections	10,189 (1,008)	13,383 (1,371)
R_{merge} (%)[b]	6.3 (32.4)	9.9 (35.63)
Completeness (%)	97.9 (98.8)	97.8 (100)
Refinement Statistics:		
Resolution range (Å)	12 - 2.6	12 - 2.6
Reflections, $F>2\sigma(F)$	9,171	13,034
Rcryst% [c]	23.3	25.5
Rfree% [d]	30.3	29.9
Number of atoms (protein/DNA/water)	1160/814/112	1149/814/103
RMS deviations:		
Bonds (Å)	0.007	.006
Angles (°)	1.1	1.1
Average B factors ($Å^2$) (protein/DNA/water)	59.8/54.1/54.2	54.5/62.7/57.4
Ramachandran Plot Quality:		
Most favored (%)	86.0	89.1
Additionally allowed (%)	13.2	10.1
Generously allowed (%)	0.8	0.8
Disallowed (%)	0	0

[a] Values for outermost shell are given in parentheses.

[b] Rmerge = $\Sigma|I - \langle I \rangle|/\Sigma|$, where I is the integrated intensity of a given reflection.

[c] Rcryst = $||Fo| - |Fc||/ \Sigma |Fo|$.

[d] Rfree was calculated using 5% (for *fkh250*) and 7.5% (for *fkh250*^{con*}) of data excluded from refinement.

Table S2. Summary of the in vitro and in vivo readouts examined for ScrWT and the three Scr mutants. This table summarizes the data shown in Figures 4, 5, and 6. Three points are apparent: 1) ScrWT is the most robust activator for all of the in vivo readouts, 2) the His-12A,Arg3A double mutant is unable to carry out any of Scr's specific functions, but can still activate the more general Hox readout, *fkh250^{con}-lacZ*, and 3) the Kd measurements for *fkh250* and *fkh250^{con}* correlate best with the activation of the *fkh250-lacZ* and *fkh250^{con}-lacZ* reporter genes, respectively. The various readouts show different sensitivities to the single point mutations, suggesting that these responses are more complex than being dependent upon a single Scr – binding site interaction. Nevertheless, these data support the view that both His-12 and Arg3 are important for Scr to carry out its specific functions in vivo. Kds are in nM and were measured in the presence of Hth^{HM}-Exd. The in vivo effects of the Gln4G mutant were not measured (ND) because, based on the results with Arg3A, they are expected to only partially affect function, at best.

protein	**Kd-250**	**T1**	**α-CrebA**	**α-Fkh**	***fkh250-lacZ***	***fkh250^{con}-lacZ***	**Kd-250^{con}**
ScrWT	8.6	++++	+++	+++	++++	++++	10.3
His-12A	14.9	++++	++	++	+++	++++	12.0
Arg3A	47.5	+++	+	–	–	++++	28.6
His-12A,Arg3A	54.6	–	–	–	–	++++	18.2
Gln4G	49.7	ND	ND	ND	ND	ND	26.6

Joshi et al, Supplementary Figure 1

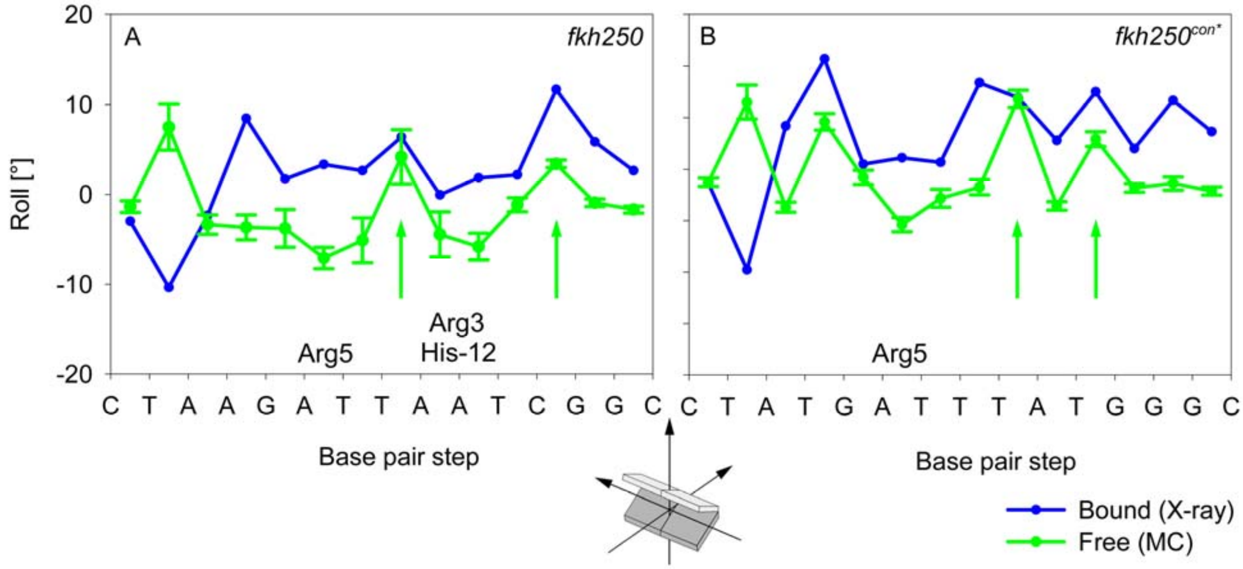

Figure S1. Base pair roll of *fkh250* and *fkh250*^{con*}.

Graphs showing roll as a function of base pair steps for *fkh250* (A) and *fkh250*^{con*} (B). The blue curves represent the roll as observed in the X-ray structures. The green curves show the roll predicted by the MC simulations. The roll pattern is distinct for each DNA, seen most readily in the spacing of the two peaks highlighted with arrows. In both cases, the MC simulations result in a similar roll pattern compared to the X-ray structures. Differences in minor groove geometry result from sequence-specific variations of helical parameters, with the variation in roll as a predominant factor in determining the minor groove geometry. A negative roll between two adjacent base pairs compresses the DNA minor groove whereas a positive roll compresses the DNA major groove and thus widens the minor groove at the location of this base pair (Rohs et al., 2005b). The regions in *fkh250* where Arg5 and Arg3/His-12 insert into the minor groove are characterized by extended roll minima (A), while *fkh250*^{con*} shows a single region with low roll values at the Arg5 position. In both structures, the large positive roll value at the central TpA step counters minor groove narrowing.

A negative roll accompanied by a positive roll half-a-helix turn away enhances minor groove narrowing at the base pair which shows the negative roll. Therefore, distances between base pair steps showing distinctive roll maxima and minima are determinants of groove geometry (Rohs et al., 2005b). Here, the more distinct minor groove narrowing of the *fkh250* DNA is assisted by the positive roll of the CpG step almost half-a-helix turn away from the center of the AT-rich region. The TpA and TpG steps with roll maxima in the *fkh250*^{con*} DNA are in close proximity, thus widening the minor groove. TpA steps usually show a positive roll and can be affected by binding and crystal packing due to their unfavorable stacking, which is reflected by the 5'-end TpA step.

Joshi et al, Supplementary Figure 2

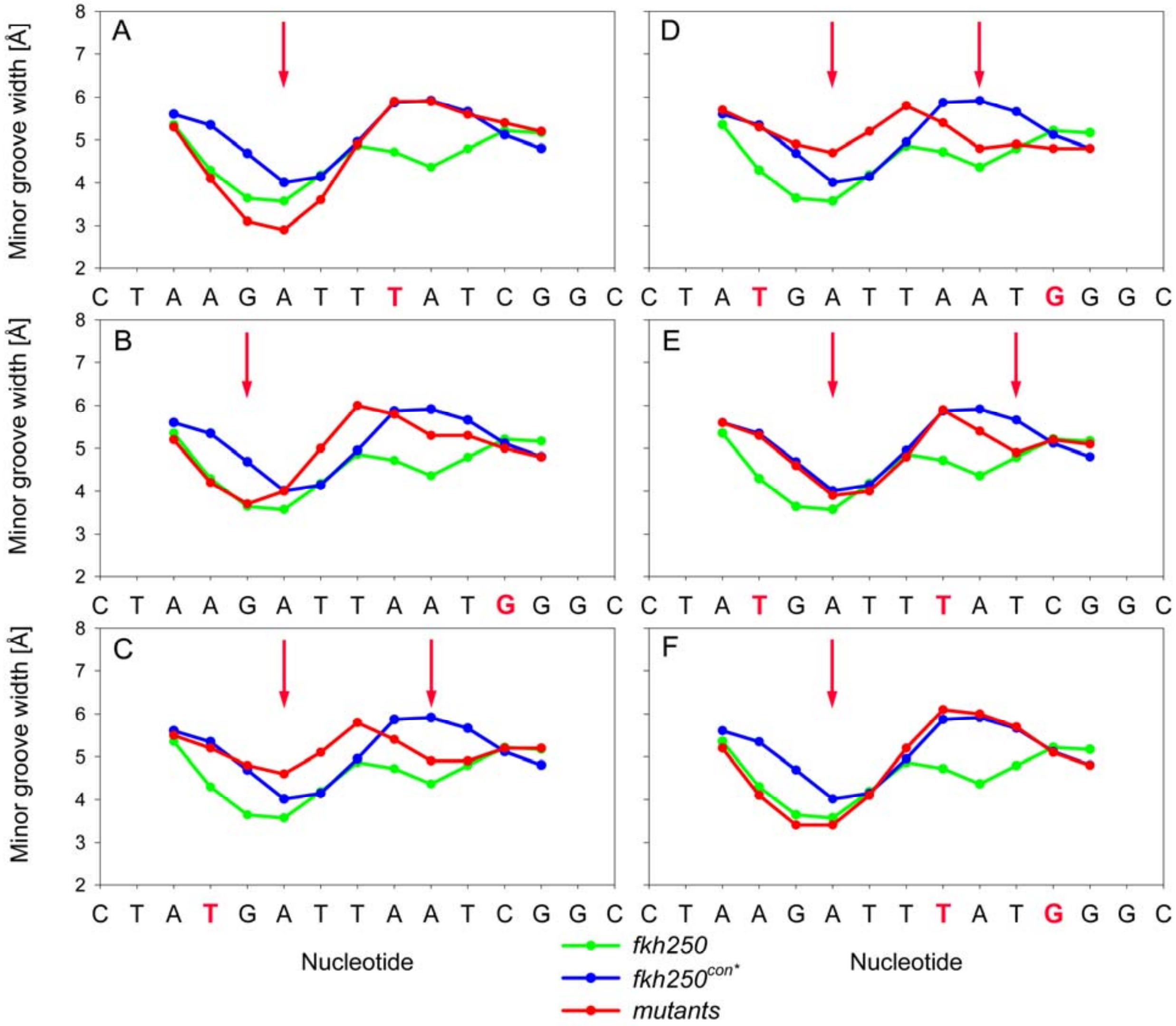

Figure S2. MC analysis of minor groove width for DNAs with individual base pair differences.

fkh250 and *fkh250^{con*}* have three base pair differences (see Figure 1B). In this figure, we use the MC method to analyze the effects on minor groove width of all possible combinations of these three differences.

A, B, C: MC simulations of single base pair changes, relative to *fkh250*. The three base pairs in which the *fkh250* sequence differs from the *fkh250^{con*}* sequence were separately replaced by the respective base pair of the *fkh250^{con*}* DNA. Mutants are shown in red in comparison to the *fkh250* (green) and *fkh250^{con*}* (blue) MC simulations. Mutation of the central A-T base pair (A) and the 3'-C-G base pair (B) open the minor groove mainly in the region where Arg3 and His-12 bind. The single mutation of the 5'-A-T base pair (C) raises the absolute values of the overall groove width. Minor groove width minima in the MC simulations of the mutants are highlighted by arrows.

D, E, F: MC simulations of double base pair changes, relative to *fkh250*. Mutants are shown in red in comparison to the *fkh250* (green) and *fkh250*^{con*} (blue) MC simulations. The double mutant without mutating the central A-T step (D) raised the absolute values of the overall groove width. Mutating the 3'-A-T and central A-T base pairs simultaneously (E) conserved two minima in groove width but also raised absolute values in the region where Arg3 and His-12 bind. The *fkh250* double mutant with both the central A-T and 3'-C-G base pair mutated closely resembled the *fkh250*^{con*} minor groove geometry (F). The sequence analyzed in this panel (F) is identical to *fkh250*^{con} (Ryoo and Mann, 1999). These data indicate that the central A-T base pair difference has the largest effect, followed by the 3'-C/G difference, and to a smaller degree by the 5'-A/T difference. Minor groove width minima are highlighted by arrows.

Joshi et al, Supplementary Figure 3

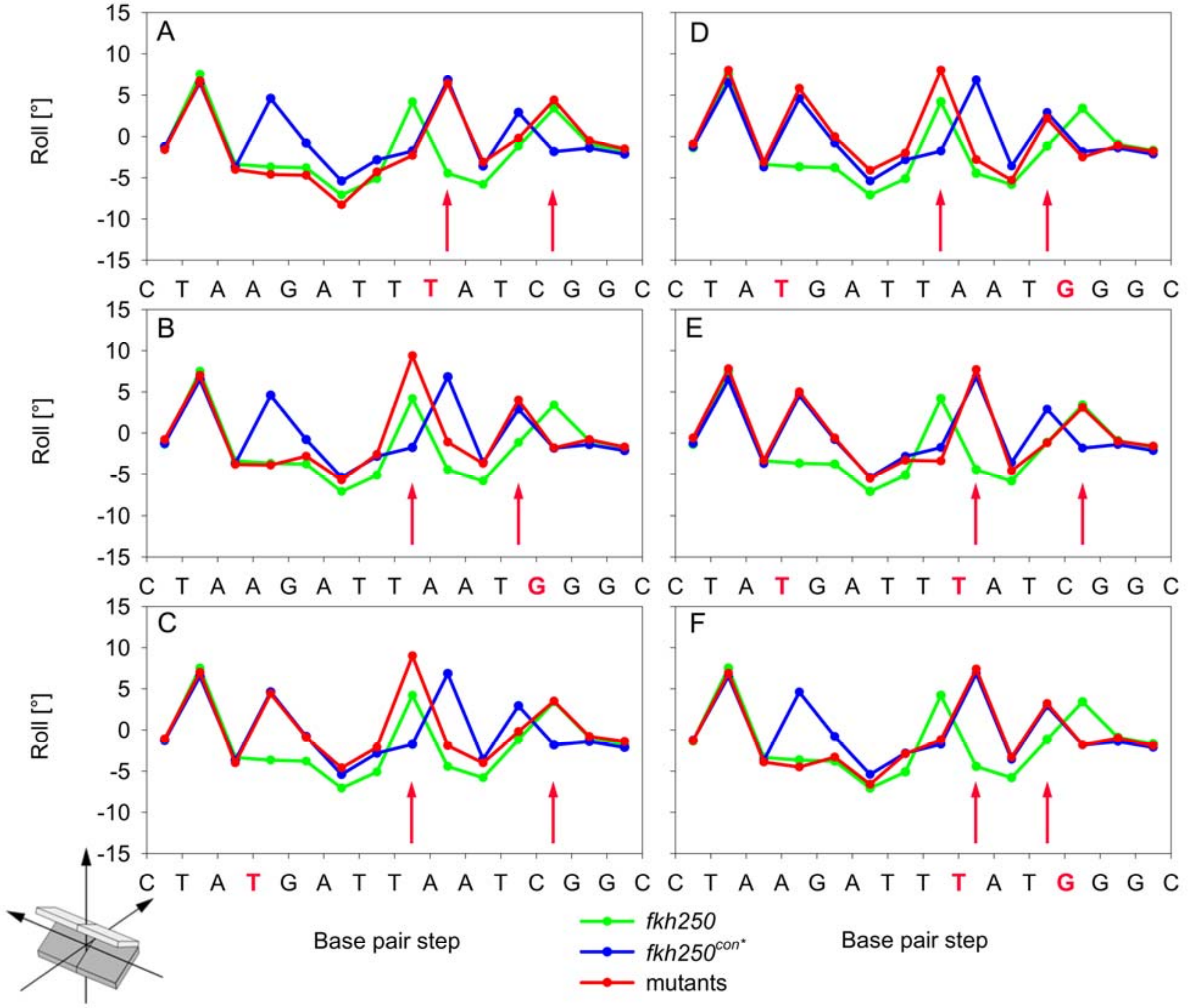

Figure S3. MC analysis of roll for DNAs with individual base pair differences.

fkh250 and *fkh250*^{con*} have three base pair differences (see Figure 1B). In this figure, we use the MC method to analyze the effects on roll of all possible combinations of these three differences.

A, B, C: MC simulations of roll for single base pair differences (color code as in Supplementary Figure 2). The mutation of the central A-T base pair shifted the roll maximum by one base pair in 3'-direction, where *fkh250*^{con*} shows a roll maximum (A). In comparison, mutation of the 3'-C-G base pair shifted the roll maximum by one base pair in 5'-direction, where *fkh250*^{con*} also shows a roll maximum (B). Mutating the 5'-A-T base pair generated an additional roll maximum (C) explaining the increased minor groove width without changing the pattern upon this mutation. The two peaks in the 3' region are highlighted by arrows.

D, E, F: MC simulations of roll for double base pair differences (color code as in Supplementary Figure 2). The effect of double mutations can be localized in the roll pattern. In agreement with the single mutants, simultaneous mutations of the 5'-A-T and 3'-C-G base pairs generated an additional 5’-roll maximum and shifted the 3’-roll maximum by one base pair in 5’-direction (D). Similarly, the double mutant involving both A-T base pairs showed an additional 5’-roll maximum and shifted the central roll maximum by one base pair in 3’-direction (E). Mutating the central A-T and 3’-C-G base pairs simultaneously (F) brought the two roll maxima, which are four base pairs away from each other in the *fkh250* DNA, into close proximity resembling closely the *fkh250*^{con*} roll pattern with the two maxima only separated by one base pair step. The sequence analyzed in this panel (F) is identical to *fkh250*^{con} (Ryoo and Mann, 1999). Modifying the distance between these two roll maxima affects the minor groove geometry most dramatically. The two peaks in the 3’ region are highlighted by arrows.

In general, all mutations exchange the order of pyrimidine (Y) and purine (R) bases and thus shift the location of YpR base pair steps along the sequence. The location of these steps is crucial to their either enhancing or countering the effect on minor groove narrowing. Our observation of larger roll values of YpR base pair steps in comparison to RpR and RpY base pair steps for the specific Hox binding sequences is found to resemble a general tendency in crystal structures of protein-DNA complexes (Olson et al., 1998).

Joshi et. al., Supp. Figure 4

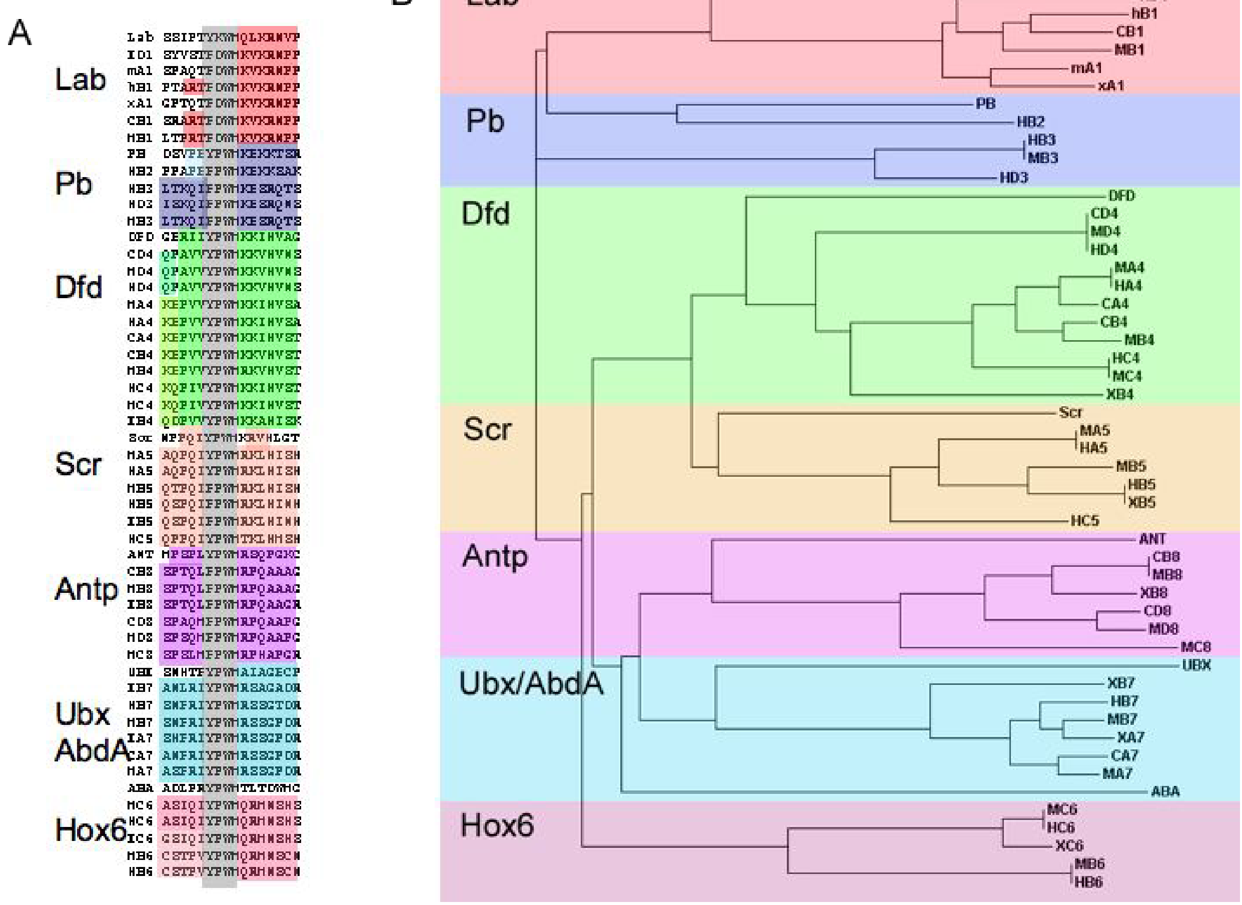

Figure S4. Hox paralogs cluster according to sequences close to the YPWM motif.

A: ClustalW-based alignments of sequences surrounding the YPWM motifs (-5 to +7) of a wide range of Hox proteins from multiple species. The Hox paralogs (labeled on the left according to the Drosophila gene names) cluster together, with the exception of vertebrate Hox6 orthologs, which, interestingly, are grouped separately. AbdB orthologs were not included in this analysis because they do not share a YPWM motif. Abbreviations: Lab: Labial; PB: Proboscipedia; DFD: Deformed; Scr: Sex combs reduced; ANT: Antennapedia; UBX: Ultrabithorax; ABA: Abdominal-A; X: Xenopus; m: mouse; C: Chicken; H: Human.

B: Tree-based diagram of the ClustalW output shown in A. Paralog-specific clustering is observed, demonstrating a clear paralog-specific relationship of Hox YPWM regions.

Supplemental References

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1996). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc *118*, 2309-2309.

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 U S A *95*, 11163-11168.

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.

Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A., and Honig, B. (2002). Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. J Comput Chem *23*, 128-137.

Rohs, R., Bloch, I., Sklenar, H., and Shakked, Z. (2005a). Molecular flexibility in ab initio drug docking to DNA: binding-site and binding-mode transitions in all-atom Monte Carlo simulations. Nucleic Acids Res *33*, 7048-7057.

Rohs, R., Sklenar, H., and Shakked, Z. (2005b). 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.

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.