Cell

Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences

Michael F. Berger,[1,3,8] Gwenael Badis,[5,8] Andrew R. Gehrke,[1,8] Shaheynoor Talukder,[5,8] Anthony A. Philippakis,[1,3,6] Lourdes Peña-Castillo,[4] Trevis M. Alleyne,[5] Sanie Mnaimneh,[4] Olga B. Botvinnik,[1,7] Esther T. Chan,[5] Faiqua Khalid,[4] Wen Zhang,[5] Daniel Newburger,[1] Savina A. Jaeger,[1] Quaid D. Morris,[4,5] Martha L. Bulyk,[1,2,3,6,*] and Timothy R. Hughes[4,5,*]
[1]Division of Genetics, Department of Medicine
[2]Department of Pathology
Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[3]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138, USA
[4]Banting and Best Department of Medical Research
[5]Department of Molecular Genetics
University of Toronto, Toronto, ON M5S 3E1, Canada
[6]Harvard/MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115, USA
[7]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[8]These authors contributed equally to this work
*Correspondence: mlbulyk@receptor.med.harvard.edu (M.L.B.), t.hughes@utoronto.ca (T.R.H.)
DOI 10.1016/j.cell.2008.05.024

SUMMARY

Most homeodomains are unique within a genome, yet many are highly conserved across vast evolutionary distances, implying strong selection on their precise DNA-binding specificities. We determined the binding preferences of the majority (168) of mouse homeodomains to all possible 8-base sequences, revealing rich and complex patterns of sequence specificity and showing that there are at least 65 distinct homeodomain DNA-binding activities. We developed a computational system that successfully predicts binding sites for homeodomain proteins as distant from mouse as *Drosophila* and *C. elegans*, and we infer full 8-mer binding profiles for the majority of known animal homeodomains. Our results provide an unprecedented level of resolution in the analysis of this simple domain structure and suggest that variation in sequence recognition may be a factor in its functional diversity and evolutionary success.

INTRODUCTION

The approximately 60 amino acid homeobox domain or "homeodomain" is a conserved DNA-binding protein domain best known for its role in transcription regulation during vertebrate development. The homeodomain can both bind DNA and mediate protein-protein interactions (Wolberger, 1996); however, the precise mechanisms that dictate the physiological function and target range of individual homeodomain proteins are in general either unknown or incompletely delineated (Banerjee-Basu et al., 2003; Svingen and Tonissen, 2006). In several cases, functional specificity can be traced to the homeodomain itself (Chan and Mann, 1993; Furukubo-Tokunaga et al., 1993; Lin and McGinnis, 1992), indicating that individual homeodomains have distinct protein- and/or DNA-binding activities. Since many homeodomains have similar DNA sequence preferences, much attention has been paid to the role of protein-protein interactions in target definition (Svingen and Tonissen, 2006), despite evidence that the sequence specificity of monomers contributes to targeting specificity (Ekker et al., 1992) and that binding sequences do vary, particularly among different subtypes (Banerjee-Basu et al., 2003; Ekker et al., 1994; Sandelin et al., 2004). Indeed, it has been proposed that the DNA-binding specificity of homeodomains is determined by a combinatorial molecular code among the DNA-contacting residues (Damante et al., 1996).

Efforts to understand the physiological and biochemical functions of homeodomains have been hindered by the fact that most have only a few known binding sequences, if any. Position weight matrices (PWMs) have been compiled for 63 distinct homeodomain-containing proteins from human, mouse, *D. melanogaster*, and *S. cerevisiae* in the JASPAR (Bryne et al., 2008) and TRANSFAC (Matys et al., 2003) databases. These matrices are based on 5 to 138 individual sequences (median 18), presumably capturing only a subset of the permissible range of binding sites for these factors. Further, the accuracy of PWM models has been questioned (Benos et al., 2002), and there are many examples in which transcription factors bind sets of sequences that cannot be described in a conventional PWM representation (Blackwell et al., 1993; Chen and Schwartz, 1995; Overdier et al., 1994).

Moreover, the sequence preferences of the individual proteins can, in some cases, be altered by the binding context: For instance, the binding specificity of the complex of *Drosophila* Hox-Exd homeodomain proteins is remarkably different from

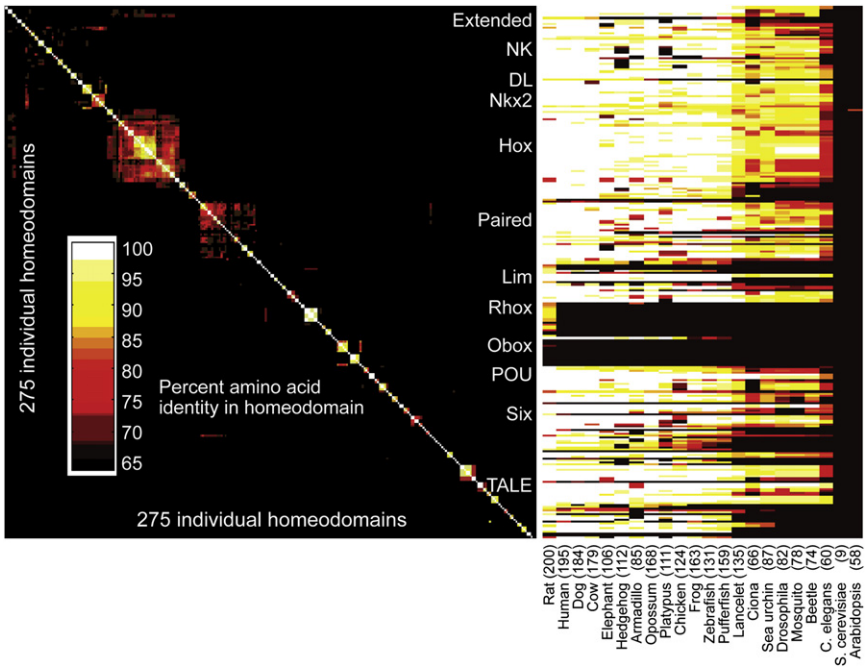

Figure 1. Conservation and Diversity of Mouse Homeodomains

Left: Heat-map showing the percent identity between different hierarchically clustered mouse homeodomains. Major homeodomain families are indicated. Right: percent identity to closest BLAST or BLAT hit in other species as indicated. The number of distinct homeodomain-containing protein counterparts in other species is given at bottom (isoforms are counted as a single entity).

that of the individual monomers (Joshi et al., 2007), raising the prospect that the monomeric binding preferences may not always be relevant to targeting in vivo. There is evidence that the sequence preferences of individual Hox proteins in *Drosophila* and mammals are significantly altered by physical interactions with protein cofactors in the PBC and Meis subfamilies, presumably through contacts to the Hox N-terminal arm that change the way the homeodomain contacts DNA (Mann and Chan, 1996; Wilson and Desplan, 1999). Other evidence, however, suggests that these examples of cofactor alterations to the monomer binding specificities are likely to be the exception rather than the rule. Carr and Biggin demonstrated that there is good correlation between monomer binding in vitro and in vivo for four fly homeodomain-containing proteins: Eve, Ftz, Bcd, and Prd (Carr and Biggin, 1999). Carroll and colleagues further showed that Ubx activity in promoting haltere development is independent of protein cofactors and that the promoters of its target genes in this pathway contain clusters of individual Ubx binding sites (Galant et al., 2002). Liberzon et al. showed not only that the specificity of the Hox-like mouse protein Pdx1 also extends beyond the TAAT core, but that the preferences at these flanking positions in vitro correlate with the ability of these sequences to stimulate transcription in vivo (Liberzon et al., 2004). In addition, for many domain classes, and in organisms ranging from yeast to human, in vivo binding sites detected by ChIP-chip typically contain sequences that reflect those preferred in vitro (Carroll et al., 2005; Harbison et al., 2004).

The mouse genome encodes a larger number of homeodomains than most vertebrates, including humans, and contains representatives of both ancient (NK, Hox) and young (Rhox, Obox) homeodomain families, encompassing striking examples of both purifying and diversifying selection (Jackson et al., 2006; Larroux et al., 2007; Rajkovic et al., 2002). The mouse homeodomain complement, estimated at 260 distinct proteins and 275 individual homeodomains (Bult et al., 2004), is broadly conserved across animals (Figure 1). For example, most mouse homeodomains (172/275, or 63%) have an identical human counterpart, and among these, most (107/172) have fewer than ten amino acid differences from their *Drosophila* counterpart. In contrast to their relative invariance over evolutionary time, however, most homeodomains within a genome are very different from other homeodomains within the same genome (Figure 1): Although there are 22 instances of mouse proteins with identical homeodomains, the median number of amino acid differences between any two mouse homeodomains is 37.

In this analysis, we sought to fully characterize the sequence preferences of mouse homeodomains in order to ask whether the binding activity is unique to each homeodomain and whether the full activity profile can be predicted from the primary amino acid sequence of the homeodomain, in a way consistent with a molecular code. We also explore the relevance of the monomeric binding preferences to binding sites in vivo. Since the mouse homeodomains exemplify the functional diversity inherited from the common ancestor of all animals, as well as the potential for homeodomain expansion and divergence, our results and conclusions are extendible across the animal kingdom.

RESULTS

Analysis of the Binding Preferences of Mouse Homeodomains to All 8-mers

Structures of homeodomains binding to DNA, as well as in vivo and in vitro selected binding sequences, are consistent with a typical binding footprint of seven or eight bases for a homeodomain monomer (Banerjee-Basu et al., 2003; Sandelin et al., 2004). To analyze the DNA-binding specificity, we used protein binding microarrays (PBMs) (Mukherjee et al., 2004) containing 41,944 60-mer probes in which all possible 10-base sequences are represented. Moreover, all nonpalindromic 8-mers occur on at least 32 spots on our microarray in different sequence contexts, thus providing a robust estimate of the binding preference of each protein to all 8-mers (Berger et al., 2006). For the facilitation of inference of wider motifs, the arrays also contain 32 instances of all gapped 8-mers up to a width of 12 bases. In total, we can reliably derive quantitative binding data for 22.3 million

gapped and contiguous 8-mers (4^8 sequence variants of 341 patterns up to 8 of 12) for any single protein. We used PBMs to analyze 194 of the 260 mouse homeodomain proteins for which we were able to produce protein as T7-driven, GST-tagged constructs by either in vitro transcription and translation or expression and purification from *E. coli*.

We systematically quantified the relative preference of each homeodomain for all possible 8-mers by several measures. These data, together with the raw microarray intensities, are in the Supplemental Data available online. The median normalized signal intensity from each 8-mer (and its Z score transform) scale almost linearly with K_a, when known (Berger et al., 2006), but may be sensitive to the amount of protein used in the assay (data not shown). We can additionally express the binding specificity of each protein as a mononucleotide PWM, or motif (contained in Table S1), but these often fail to fully capture the complete spectrum of binding activities and lack the resolution provided by individual word-by-word measurements (Benos et al., 2002; Chen et al., 2007). Here, we primarily employ a statistic we refer to as the enrichment score (E score) for each 8-mer, which is a variation on area under the ROC curve (AUC) and scales from 0.5 (highest) to −0.5 (lowest) (Berger et al., 2006). This measure is unitless and has a nonlinear scaling with intensity (there is a compression of the dynamic range among the most highly bound sequences), but on the basis of rank correlations and precision-recall analysis it is the most highly reproducible of any measure we have tested (Figure S3), and it facilitates comparison between separate experiments. On the basis of random permutations of the array data, our entire data set should contain no randomly arising E scores above 0.45. Using E > 0.45 for at least one 8-mer as a PBM success criterion, we obtained clear sequence preferences for 168 homeodomain proteins, including 11 different factors with identical homeodomain amino acid sequences. On average, each homeodomain had 144 such ungapped preferred 8-mers. It is possible that some proteins for which no sequence preference was obtained were improperly folded. The 26 we scored as unsuccessful, however, include seven of the nine Rhox isoforms tested, all three of the Lass isoforms tested, and both Satb isoforms tested, suggesting that these classes bind DNA nonspecifically or not at all or require modifications or cofactors not present in these experiments. This conclusion is supported by previous observations that *S*pecial *A*-*T*-rich *b*inding protein 1 (Satb1) binding preferences relate primarily to nucleotide composition and not to a specific sequence (Dickinson et al., 1992), a trend which is also present in our data (data not shown). Each of these 12 proteins exhibits a nonconsensus amino acid in at least one of the four positions conserved across nearly all homeodomains (positions 48, 49, 51, and 53 [Banerjee-Basu et al., 2003]), as do the majority of all failures that we obtained. Nonetheless, we observed sequence-specific binding for nine nonconsensus homeodomains, including Rhox6 and two novel homeodomains we have termed Dobox4 and Dobox5, indicating a potential means for acquiring additional diversity in DNA-binding specificity and function.

Comparison of PBM Data to Previously Determined Homeodomain Binding Preferences

As a first step in the analysis of our data, we compared our data to previously known binding sequences from the literature. Taking the 168 mouse proteins together with their closest ortholog in other metazoan species (regardless of the degree of similarity), the TRANSFAC and JASPAR databases contain at least one binding sequence corresponding to 97 mouse proteins or their orthologs (see the Supplemental Data for details). None of these proteins has more than 86 known binding sites, either in vitro or in vivo, in these databases. Nine of them (or an ortholog) have a PWM in the JASPAR database (derived from between 10 and 59 sequences obtained in vivo, in vitro, or both), and 58 more (or an ortholog) have a PWM in TRANSFAC (derived from between 5 and 86 binding sequences). An additional 30 of the 168 proteins we analyzed have between one and four known sites listed with a direct interaction observed in vivo or in vitro. We note that there are frequently multiple mouse homologs for each homeodomain in other species (e.g., Antp is the closest *Drosophila* homolog to the mouse Hox6, Hox7, Hox8, and Hox9 paralogs, so the Antp PWM represents the only data available for nine of the mouse homeodomains we analyzed).

Although the accuracy of the standard PWM model has been called into question, PWMs represent a straightforward means to compare binding activities on a coarse level. A visual comparison of the PWMs we derived from our data and those in the databases reveals reassuring similarities but also discrepancies with the existing literature (Table S1). For example, our PWMs for Lhx3, Meis1, Otx1/2, Nkx2-2, Pitx2, and Tgif1 are very similar to those previously determined. In some cases, however, our PWMs are somewhat different; for example, our Hmx3 PWM (resembling CAATTAA) is different from that previously determined from nine in vitro selected DNA sequences (resembling CAAGTGCGTG), although ours is very similar to those we obtained for the related proteins Hmx1 and Hmx2.

Perhaps the most obvious source of disagreement would be inconsistency in the initial data used to construct the motifs. We compared whether the individual sequences from JASPAR, which are determined by curators to be high quality, all contain 8-mers with high scores in our data. In some cases, all of the source sequences in JASPAR contain at least one 8-mer with an E score $\geq$ 0.45 in our data for the same protein; for example, all 41 of the human and mouse Lhx3 binding sequences meet this criterion, as do 17/18 Pbx1 binding sequences and 32/38 Nobox (Og2x) binding sequences. All of these proteins also have a PWM that is very similar to the one we derived from our data. In contrast, only one of ten in vitro selected sequences for the mouse En1 protein contains an 8-mer with E > 0.45 in our data, and the derived PWMs bear little resemblance (Table S1). Notably, the measured binding affinity of En1 for this one sequence was considerably higher than for any of the other nine selected sequences (Catron et al., 1993).

We conclude that our data are in most cases consistent with previous data, although in many cases there are discrepancies. We note that the previous data are also not always in agreement with each other; for example the En1 PWMs in TRANSFAC and JASPAR are quite different from each other, and also from the *Drosophila* Engrailed PWM in TRANSFAC, illustrating that motifs in databases and the literature cannot all be taken as a gold standard. We propose that heterogeneity in methods used to produce the DNA-binding data in the literature may underlie many of the differences between our results and previous findings:

Not only were the binding sites for separate proteins identified by different means, but even individual TRANSFAC matrices for single proteins are frequently derived from binding sequences compiled from multiple experimental methods. Further, these sequences often exhibit ascertainment bias reflecting which particular sequences were chosen to be examined by the investigators. In contrast, our data are homogeneous and were generated on a uniform, unbiased platform under standardized conditions, such that the binding activities of the different proteins should be directly comparable.

For 71 of the proteins we analyzed, there is no in vitro or in vivo binding site data, and for the majority, there is no PWM, in either mouse or the closest homolog in any species. To our knowledge, for several families, we describe a relatively uniform and apparently distinct binding profile for the first time. These encompass the Irx family (preferring sequences resembling TACATGTA), the Obox family (GGGGATTA), the Six family [G(G/A)TATCA], Gbx1/2 (CTAATTAG), and Pknox1/2 (CCTGTCA). Our data also include individual proteins with apparently unique sequence preferences, including Dux1 (CAATCAA), Hdx [(C/A)AATCA], Hmbox (TAACTAG), Homez (ATCGTTT), and Rhox11 [GCTGT(T/A)(T/A)]. The variety in motifs we obtained motivated us to further explore the similarities and differences among homeodomains within our data set.

Homeodomains Have Rich and Diverse Sequence Preferences

Figure 2A shows a 2D clustering analysis of the E scores of all 2585 8-mers that were bound by at least one homeodomain with E > 0.45. On a coarse level, the major features of the data structure correspond to the major homeodomain subclasses, and these large clusters contain sequences similar to those previously established for these subclasses, when known (Banerjee-Basu et al., 2003). For example, the largest feature (encompassing the upper-left part of Figure 2A) includes the Hox subclasses and other homeodomains that prefer a canonical TAAT core (Svingen and Tonissen, 2006). Roughly half of the homeodomains, however, have a stronger preference for other sequences, and many of the homeodomains that do bind canonical sequences also bind additional sequences (e.g., some of the Lhx classes are associated with the large TAAT binding cluster, but also have their own clusters of preferred 8-mers, boxed in Figure 2A). There are also instances of single proteins or small groups that have a distinctive 8-mer profile (Figure 2A). Indeed, when considering the top 100 highest-affinity 8-mers for each homeodomain, we identified 33 clearly separate DNA-binding activities. These binding profiles are distinguishable on the basis of limited overlap among the top 100 8-mers (among all 32,896 possible 8-mers when reverse complements are merged) for pairs of homeodomains (Figure 2B). As controls, our dataset includes 21 instances in which the same homeodomain was analyzed twice, either (1) as a freshly expressed aliquot from the same construct (three proteins) or an alternate construct (seven proteins) or (2) as a different gene with the same homeodomain sequence but different flanking residues (11 proteins). These 21 replicates invariably correlate highly: Among them, the top 100 overlap was 85 ± 8, such that proteins sharing fewer than 66 of 100 top 8-mers (99% confidence interval) were considered to have distinct binding activities. Figure 2B shows the resulting 33 specificity groups along the diagonal, accompanied by PWMs for representative members of each of the large families.

Members within each of these 33 groups, however, can be further distinguished by their lower-affinity binding sites and/or by differences in relative preference among the top 100 8-mers. For example, among the large group in the upper left of Figure 2B (bracketed) comprised of 42 proteins that are indistinguishable by the top 100 criterion, we identified 15 distinct subgroups on the basis of differences in their E score profiles over all 8-mers (Figure 3). Even though all proteins in this large group exhibit essentially the same dominant motif, clear sequence patterns are associated with the 8-mers distinctively preferred by each subgroup, and these patterns correlate with differences in their amino acid sequences (Figure 3). This is further illustrated in Figure 4. Lhx2 and Lhx4 both bind the same highest-affinity sites (8-mers containing TAATTA) but show clear, consistent preferences for different moderate- (TAATGA versus TAATCA) and lower- (TAACGA versus TAATCT) affinity sites (Figure 4A). Lhx3 and Lhx4 show greater similarity, both in binding profile and amino acid sequence, yet they have subtly different preferences for weaker 8-mers (Figure 4B). These differences only become apparent due to the richness of our dataset in capturing precise binding specificities at word-by-word resolution.

We repeated the analysis of Figure 3 for all 18 of the major groups shown along the diagonal in Figure 2B to examine whether they could be further divided by fine-grained differences in specificity (Figure S7). We considered (1) whether the motif(s) derived for any two proteins were clearly distinct and (2) whether differences in the E score profiles between proteins also contain motifs that distinguish the two binding activities. Our analysis identified a total of 65 distinct binding patterns that have a striking correlation with amino acid sequence similarity among the homeodomains (Figure S7 and see below). Although an approximation, this likely represents a lower bound on the true number of distinct patterns; for instance, our analysis places Lhx3 and Lhx4 in the same subgroup, yet we can still discern subtle differences in their 8-mer binding profiles (Figures 3 and 4).

From this analysis, we conclude that homeodomains encode distinctive DNA-binding activities and that there are often major differences between the activities of individual proteins with similar dominant sequence preferences. We also find that the dominant motif is usually unable to explain all of the data and is inferior to the full 8-mer profile in predicting the outcome of a similar experiment on an independent array (Figure S3) (Chen et al., 2007). Rather, our results are consistent with a model in which homeodomain sequence preferences may be best described as a composite of binding activities, possibly representing different binding modes with different relative affinities. This idea is supported by the report that Nkx2-5 has two distinct binding activities, one with higher affinity than the other (Chen and Schwartz, 1995); indeed, the Nkx2 group, like Lhx3 and Lhx4, is one of the 65 groups that appears as if it may be further subdivided (Figure S7).

Moreover, even the dominant motifs we obtain do not correspond perfectly with the identities of the canonical homeodomain specificity residues. The homeodomain binds DNA predominantly through interactions between helix 3 (recognition

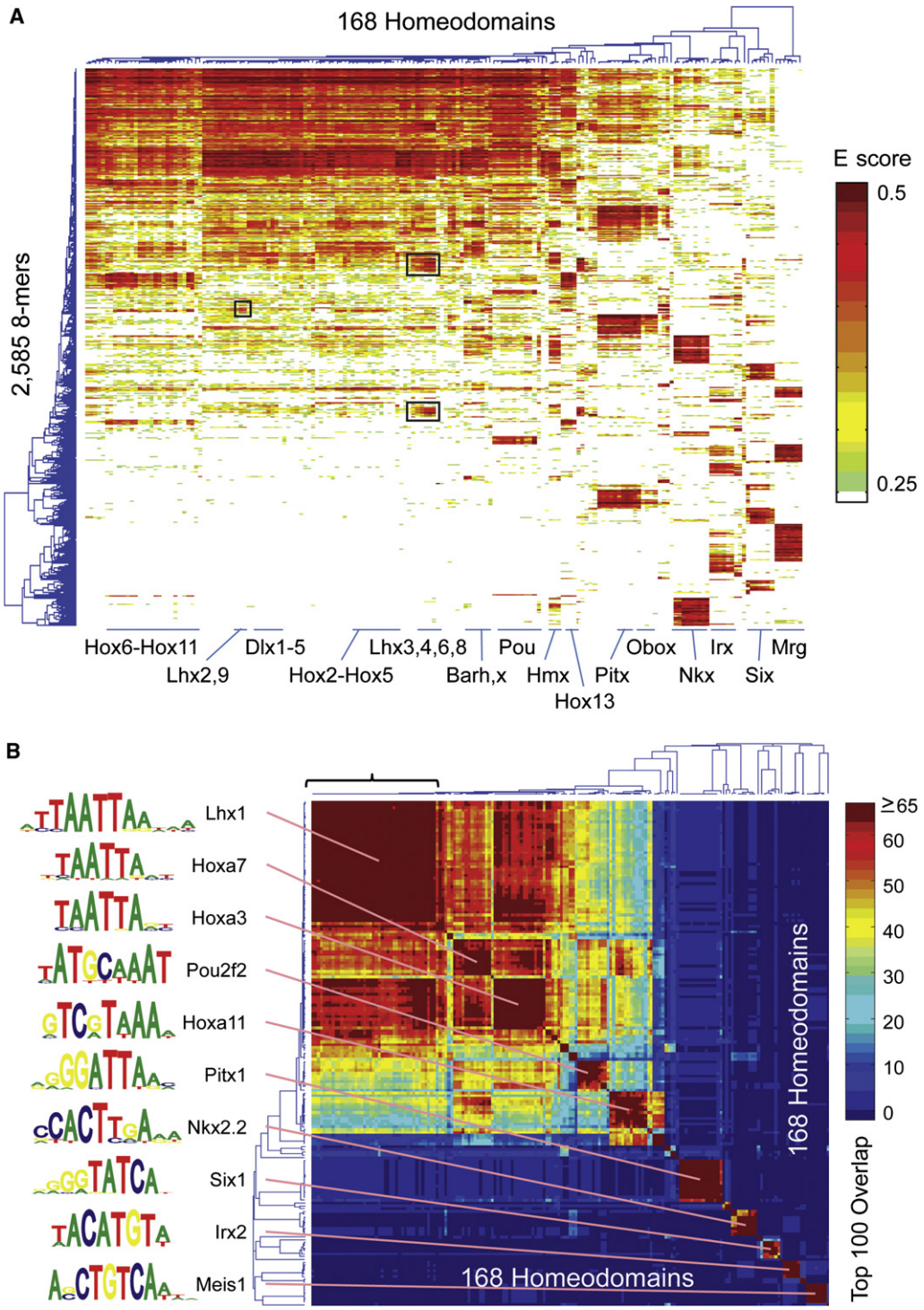

Figure 2. Overview of Homeodomains 8-mer Binding Profiles Reveals Distinct Sequence Preferences

(A) Hierarchical agglomerative clustering analysis of E score data for 2585 8-mers with E > 0.45 in at least one experiment. Boxed regions are referred to in the text. The position of exemplary homeodomain families within the dendrogram is indicated in order to highlight the diversity of overall 8-mer profiles.

(B) Clustering analysis of the matrix of overlaps in the top 100 8-mers (of all 32,896) for each pair of homeodomains. The bracket indicates the experiments analyzed in Figure 3. Logos for representative members of the major groups were determined with the Seed-and-Wobble method (Berger et al., 2006).

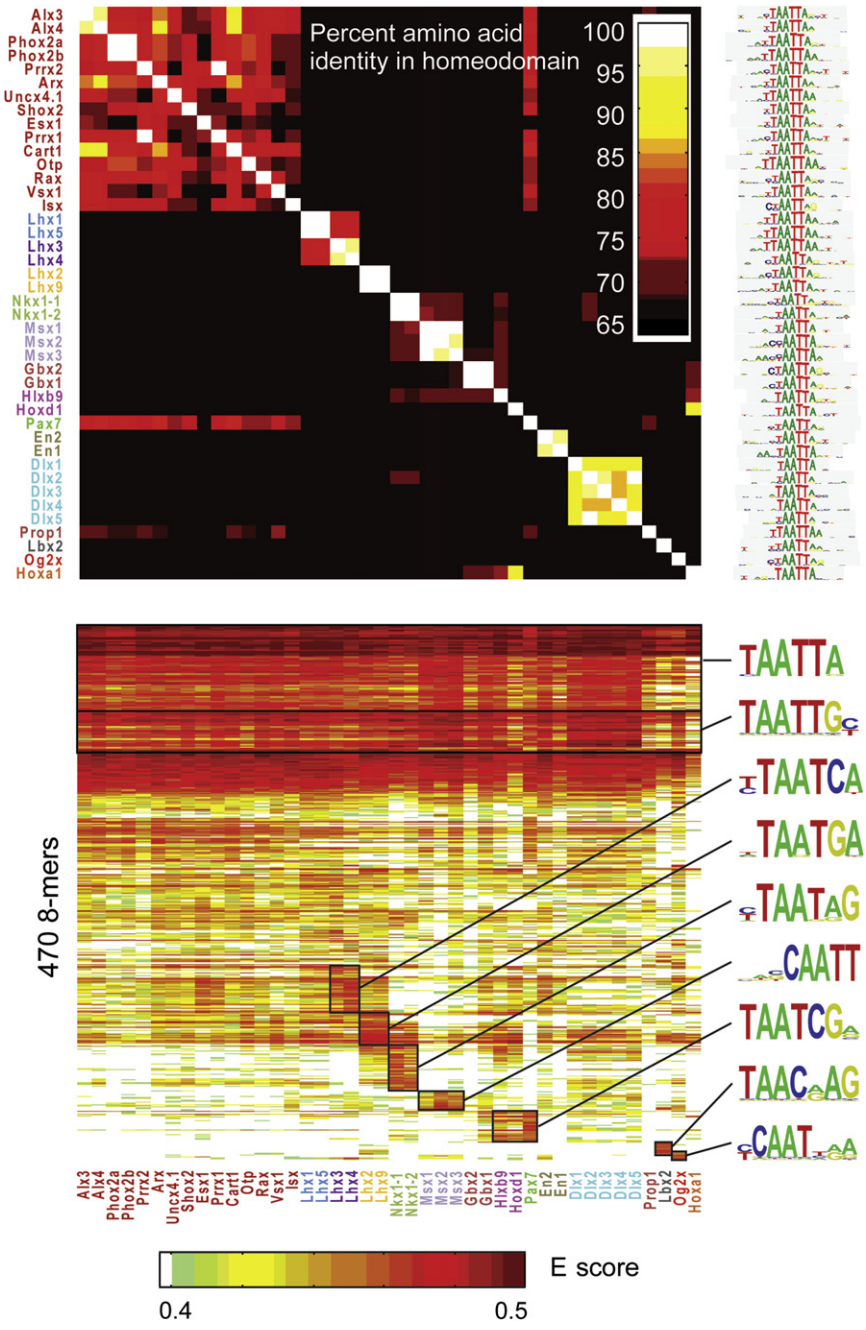

Figure 3. Homeodomains with Virtually Identical Dominant Motifs and Top 100 8-mer Preferences Have Differing Preferences for Many 8-mers

Bottom: Heat-map as in Figure 2, but restricted to the 470 8-mers with E > 0.45 in at least one of the experiments shown. Color of labels indicates groups that are distinct by our criteria. Logos were derived with ClustalW with the 8-mers in the boxed regions as inputs. Top: Amino acid similarities among these 42 homeodomains, as in Figure 1.

helix) and the major groove, and base-specific contacts made by positions 47, 50, and 54 are believed to be the main determinants of differences in binding specificity (Laughon, 1991) (Figure 5A, shown in red). Indeed, we were able to form groups harboring similar dominant motifs simply by partitioning homeodomains according to their amino acid identity at these three positions (Figure 5B). Our results are consistent with previous reports; for instance, replacement of glutamine with lysine at position 50 has been shown to dramatically alter the binding specificity through several newly formed hydrogen bonds to guanines (Tucker-Kellogg et al., 1997). These three residues alone are not sufficient to fully capture the entire binding activity, however, and in some cases, even the dominant motifs differ among proteins that have the same identity at these three residues (Figure 5B). Specific residues in the N-terminal arm have also been shown to influence binding specificities of homeodomains through minor groove interactions (Ekker et al., 1994); however, the identities at these residues (3, 6, and 7) do not correspond to the variation in Figure 5B (data not shown). Additional recognition positions must also be necessary to explain the differences in binding specificity we have observed for related homeodomains: Although we cannot exclude a molecular code controlling homeodomain DNA-binding activity (Damante et al., 1996), such a code is likely to be complex if one considers the full range of binding sequences.

Prediction of Binding Sequences across the Animal Kingdom Using Homeodomain Amino Acid Sequence Similarity

To more systematically and thoroughly approach the problem of identifying determinants of homeodomain sequence preferences, we tested the efficacy of a variety of methods to predict the full 8-mer binding profiles by using only the amino acid sequences as inputs (see the Supplemental Data for details). We evaluated each approach using leave-one-out crossvalidation (in which each homeodomain in turn was "held out" and its full 8-mer binding profile was predicted) to test our success at reproducing the 8-mer data for each of the 157 non-identical homeodomains, using Spearman correlation, top 100 overlap, and root mean squared error as success criteria in predicting the 8-mer profile. The most effective overall approach was a nearest-neighbor method, in which the 8-mer data were transferred from the homeodomain with the fewest number of mismatches over a set of 15 DNA-contacting amino acids

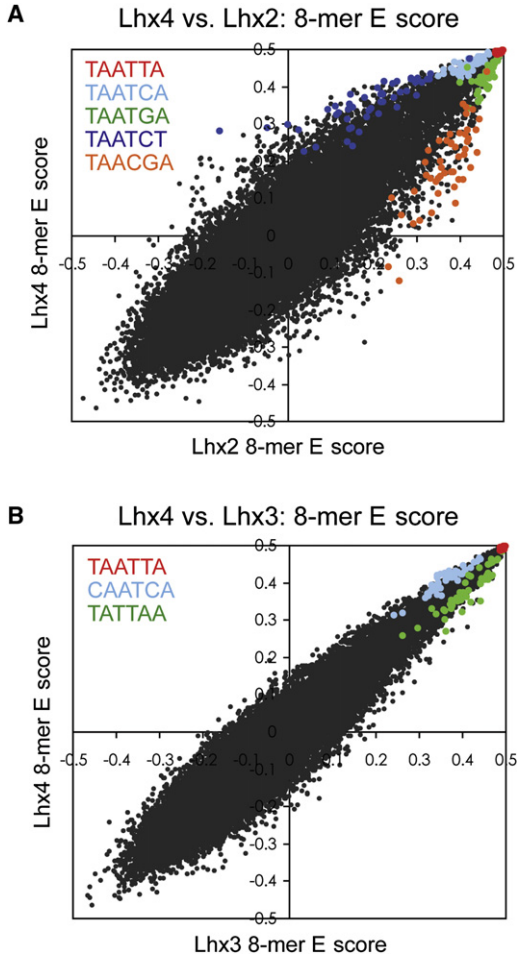

Figure 4. Scatter Plots Showing Differences in E Scores for Individual 8-mers between Lhx Family Members

(A) Comparison of Lhx2 and Lhx4.

(B) Comparison of Lhx3 and Lhx4.

8-mers containing each 6-mer sequence (inset) are highlighted, revealing clear systematic differences between sequence preferences despite essentially identical dominant motifs and sets of top 100 8-mers for these homeodomains.

(averaging the E scores in the case of ties). These 15 residues (3, 5, 6, 25, 31, 44, 46, 47, 48, 50, 51, 53, 54, 55, and 57; Figure 5A) account for all specific base-pair and phosphate backbone contacts in crystal structures for the *Engrailed* homeodomain (Fraenkel et al., 1998; Kissinger et al., 1990). The number of overlaps between the measured and predicted top 100 8-mers correlates with the distance to the closest example in the data, with zero, one, or two mismatches typically yielding predictions that are as close as an experimental replicate (Figure 6A). This result is consistent with our previous assessment of homeodomain DNA-binding activity subclassifications because there are more than 65 different naturally occurring variants among these 15 residues, groupings of which closely correspond to those obtained from the 8-mer profiles (see the Supplemental Data for details).

Consistent with the fact that much of the amino acid sequence variation among animal homeodomains is found in the mouse (Figure 1), the number of mismatches among these 15 amino acids from most mouse homeodomains to their homologs in species as distant as *Drosophila* is zero (Figures 1 and 6A and the Supplemental Data). We therefore applied the nearest-neighbor approach to project high-confidence 8-mer binding profiles for homeodomain proteins in 24 species (Supplemental Data). We found that in many cases, the predicted data were consistent with known motifs and binding sequences, even when the remainder of the homeodomain sequence had diverged considerably. We experimentally determined 8-mer E scores for the *C.elegans* homeodomain protein Ceh-22 by PBM and observed striking correlation with its predicted profile (Pearson correlation = 0.93, 78 of the top 100 overlap; Figure 6B) despite an overall difference of 11 amino acids within the homeodomain to the most similar mouse protein. Our inferred 8-mer profiles closely mirror quantitative in vitro measurements for the *Drosophila Engrailed* homeodomain, as well (Figure S8) (Damante et al., 1996).

Sequences Preferred by Homeodomains In Vitro Correspond to Sites Preferentially Bound In Vivo

Finally, we asked whether the homeodomain monomer binding preferences we identified in vitro reflect sequences preferred in vivo. Anecdotally, our highest predicted binding sequences do correspond to known in vivo binding sites. For example, in the predicted 8-mer profile for sea urchin Otx, a previously identified in vivo binding sequence (TAATCC, from the Spec2a RSR enhancer) (Mao et al., 1994), is contained in our top predicted 8-mer sequence, and, more strikingly, it is embedded in our fifth-highest predicted 8-mer sequence (TTAATCCT). At greater evolutionary distance, three of the four *Drosophila* Tinman binding sites in the minimal Hand cardiac and hematopoietic (HCH) enhancer (Han and Olson, 2005) are contained within the second (TCAAGTGG), fifth (ACCACTTA), and ninth (GCACTTAA) ranked 8-mers (the fourth overlaps the 428th ranked 8-mer [CAATTGAG], but also overlaps with a GATA binding site (Han and Olson, 2005) and may have constraints on its sequence in addition to binding Tinman).

To ask more generally whether occupied sites in vivo contain sequences preferred in vitro, we examined six ChIP-chip or ChIP-seq data sets in the literature that involved immunoprecipitation of homeodomain proteins that we analyzed, or homologs of proteins we analyzed that shared at least 14 of the 15 DNA-contacting amino acids. In all cases, we observed enrichment for monomer binding sites in the neighborhood of the bound fragments, with a peak at the center (Figure 7 and Figure S9). Figures 7A and 7B show two examples, *Drosophila* Caudal (Li et al., 2008) and human Tcf1/Hnf1 (Odom et al., 2006). For Caudal, the size of this ratio peak increased dramatically with E score cutoff, indicating that the most preferred in vitro monomer binding sequences correspond to the most enriched in vivo binding sites (cutoff E > 0.49) (Figure 7D) (51% of bound fragments have such an 8-mer, versus 17% in randomly selected fragments). For Tcf1/Hnf1, however, the majority of sequences bound in vivo do not contain the best in vitro binding sequences (E > 0.49), although most do contain at least one 8-mer with E > 0.45 (Figure 7C) (53%, versus 27% in random fragments), suggesting utilization of weaker binding sites. Similar results were obtained with PWMs (data not shown). Thus, the requirement for

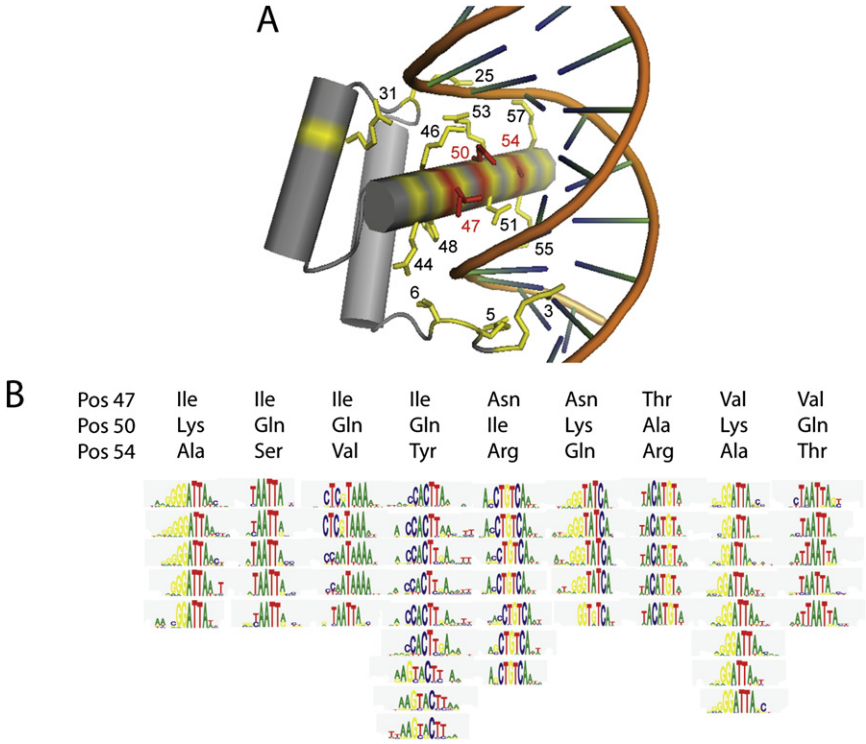

Figure 5. Correspondence between Canonical Homeodomain Amino Acid Sequence Specificity Residues and Dominant Motifs

(A) Protein-DNA interface for the *Drosophila* Engrailed protein (Kissinger et al., 1990). The three primary specificity residues discussed in the text are shown in red. The remaining residues considered in our nearest-neighbor analysis are in yellow. (B) Motifs for all homeodomains in our dataset containing each of the displayed combinations of residues. For clarity, only those combinations occurring between five and ten times are shown. Logos represent PWMs determined with the Seed-and-Wobble method (Berger et al., 2006).

highest-affinity binding sequences may vary among homeodomain proteins, species, or under different physiological contexts. Nonetheless, a large proportion of the in vivo binding events apparently involve the monomeric homeodomain sequence preferences, which can be derived in vitro.

DISCUSSION

Our data provide a new level of resolution in the analysis of homeodomain sequence specificity. Our analyses show that homeodomains have distinctive sequence preferences, which may contribute to the strong selective pressure on their amino acid sequences, as well as to the biological specificity in target genes and diversity in function among the homeodomain proteins. Our findings should provide a fertile basis for future study of homeodomain function and evolution and may influence our understanding of evolved diversity in other transcription factor families.

One of the long-standing goals in the study of DNA-protein interactions has been to elucidate the relationships between amino acid residues and base preferences. Although it is clear that key residues can exert a strong influence, with others held constant (Hanes and Brent, 1989; Treisman et al., 1989), there is also evidence that alterations in the overall structure of DNA-binding domains can influence the DNA sequence preferences in unexpected ways (Miller and Pabo, 2001; Wolfe et al., 2001). Interactions among residues in the PWM (Benos et al., 2002) further complicate derivation of a deterministic recognition code. Full 8-mer profiles provide a new way to approach this problem. Although there is a correspondence between the canonical homeodomain DNA-binding specificity residues and the dominant motif, the correspondence is imperfect, and the dominant motif does not fully describe the complete binding profile, consistent with a model in which homeodomains have multiple binding modes. Perhaps as a consequence, our analyses suggest that categorization of the 8-mer profile on the basis of the full suite of DNA-contacting residues may be a more appropriate and practical paradigm for homeodomain sequence recognition than a molecular encoding of a PWM.

This idea is supported by our accurate prediction of full binding profiles over vast evolutionary differences. In fact, it is striking how little the entire homeodomain family has evolved at DNA-contacting residues since the common ancestor of all animals, considering that the potential for diversity in homeodomain DNA-binding activity seems well suited for duplication and divergence. Although newer binding activities (e.g., those of the Oboxes, Dobox4, Dobox5, Rhox6, and Rhox11) have apparently arisen since the divergence of mice and humans (there is no apparent homolog of these homeodomains in any species more distant than rat), the range of possible configurations even at the three canonical specificity residues (47, 50, and 54) appears to be sparsely populated in nature.

In all cases we tested, including predicted profiles for *Drosophila* homeodomains, the preferred monomer binding 8-mer sequences we obtained in vitro are enriched at the center of genomic fragments bound by the same protein in vivo. From this, we conclude that monomer binding preferences are likely to be a component of targeting mechanisms in general. Other factors (e.g., the chromatin landscape and protein-protein interactions) must also play a role in targeting because only a small fraction of all possible binding sites are occupied. We cannot exclude the possibility that the homeodomains we analyzed can undergo a radical change in binding specificity when they form complexes and that they rely on this or other mechanisms for a subset of in vivo binding events. Nonetheless, our demonstration that there are strong relationships between in vitro sequence preferences and in vivo binding sites supports the biological relevance of binding preferences of homeodomain monomers and indicates that our data should be of widespread use for identifying regulatory sites in vivo.

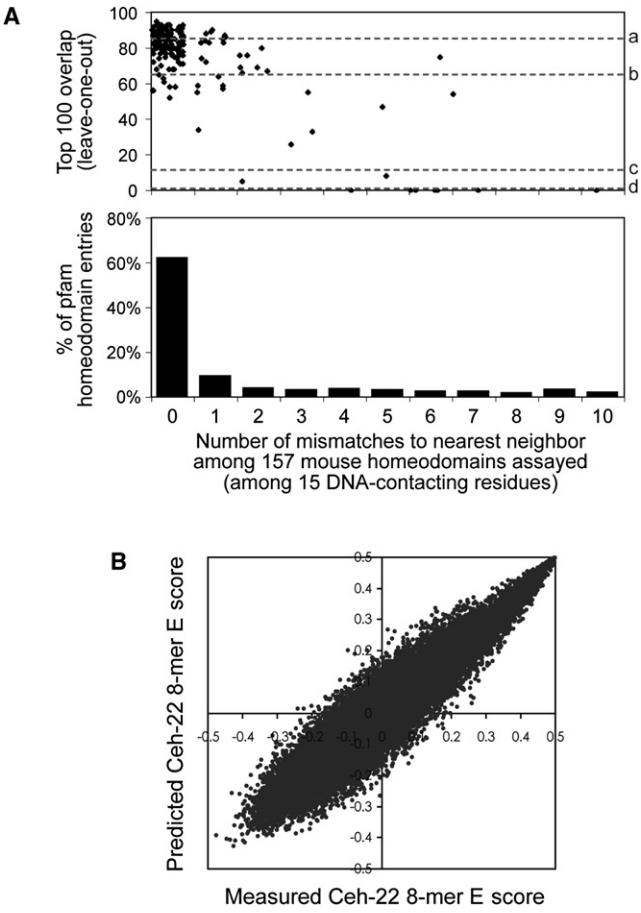

Figure 6. Correspondence between Homeodomain DNA-Contacting Amino Acid Sequence Residues and 8-mer DNA-Binding Profiles

(A) Top: Scatter plot showing the top 100 overlap between real and predicted 8-mer binding profiles from leave-one-out crossvalidation for our nearest-neighbor approach. Dashed lines indicate the following benchmarks: median, experimental replicates (a), 99% confidence, experimental replicates (b), median, randomized homeodomain labels (c) and median, randomized 8-mer labels (d). Within each bin, the x axis values have been nudged randomly for visualization. Bottom, the proportion of 3693 Pfam entries with the indicated identity to at least one mouse homeodomain analyzed.

(B) Predicted versus measured 8-mer E scores for *C. elegans* Ceh-22.

EXPERIMENTAL PROCEDURES

Cloning, Expressing, and Purifying Homeodomains

Homeodomain open reading frames, consisting of the Pfam-defined homeodomain and 15 amino acids of flanking sequence (or to the end of the full open reading frame) were cloned into pMAGIC1 (Li and Elledge, 2005) by either RT-PCR from pooled mouse mRNA or by gene synthesis (DNA 2.0). All clones were sequence verified (supplementary file "Protein and DNA sequence," available at http://hugheslab.ccbr.utoronto.ca/supplementary-data/homeodomains1/). We transferred the inserts into a T7-GST-tagged variant of pML280 following Li and Elledge (2005). We expressed proteins by either (1) purification from *E. coli* C41 DE3 cells (Lucigen) or (2) in vitro translation reactions (Ambion ActivePro Kit) without purification. Essentially identical results were obtained by either method (Figure S1).

Microarray Design and Use

The construction of "all 10-mer" universal PBMs with a de Bruijn sequence of order 10 has already been described (Berger et al., 2006) and is described in more detail in conference proceedings posted at http://thebrain.bwh.harvard.edu/RECOMB2007.pdf (Philippakis et al., 2008). For this study, we further optimized our design to achieve greater coverage of gapped *k*-mers (see the Supplemental Data for details). PBM assays were performed essentially as described previously (Berger et al., 2006), except that four proteins were simultaneously assayed in separate sectors of a single microarray and scanned with at least three different laser power settings to best capture a broad range of signal intensities and ensure signal intensities below saturation for all spots. Images were analyzed with GenePix Pro version 6.0 software (Molecular Devices), bad spots were manually flagged and removed, and data from multiple Alexa Fluor 488 scans of the same slide were combined with "masliner" software (Dudley et al., 2002) and normalized as described previously (Berger et al., 2006).

Sequence Analysis and Motif Construction

We provide several scores for each 8-mer in each experiment: (1) median intensity, (2) Z score, (3) enrichment score (E score), and (4) false discovery rate Q value for the E score. The median intensity and Z score measures follow standard statistical procedures. The E score has already been described in detail (Berger et al., 2006). In brief, for each 8-mer (contiguous or gapped), we consider the collection of all probes harboring a match as the "foreground" feature set and the remaining probes as a "background" feature set. We compare the ranks of the top half of the foreground with the ranks of the top half of the background by computing a modified form of the Wilcoxon-Mann-Whitney (WMW) statistic scaled to be invariant of foreground and background sample sizes. The E Score ranges from +0.5 (most favored) to −0.5 (most disfavored). We compute a false discovery rate Q value for the E score by comparing it to the null distribution of E scores (over 32,896 8-mers) calculated by randomly shuffling the mapping among the 41,944 probe sequences and intensities (repeated 20 times) (Subramanian et al., 2005). In computing all of the above scores, we do not consider probes for which the 8-mer occupies the most distal position on the probe (5′ with respect to the template strand) or for which the 8-mer overlaps the 24 nt primer region. We derive PWMs with the "Seed-and-Wobble" algorithm (Berger et al., 2006).

Predicting 8-mer Profiles and Scoring the Predictions

We considered two general methods for predicting 8-mer binding profiles on the basis of the primary amino acid sequence: nearest neighbor and regression. In the nearest neighbor (NN) approach, we predicted the 8-mer profile of any given homeodomain protein by taking the 8-mer profile(s) of its nearest neighbor(s) (averaging in the case of a tie). For regression, we converted the homeodomain amino acid sequence alignment to a binary representation by replacing all 20 standard amino acids in any of the canonical residue positions with unique 20 bit binary flags, the dimensionality was reduced by Principal Components Analysis (PCA), and a distinct model was learned for each 8-mer and for each homeodomain (i.e., a separate model for all 157 × 32,896 entries in the data table). We considered several variations of the distance metric used (e.g., number of mismatches versus amino acid similarity scores) and/or the residues considered (all 57 residues, 15 DNA-contacting residues, or five known specificity residues).

Chromatin Immunoprecipitation Analyses

We obtained 1331 bound sequences in the Caudal data set by selecting those in the 1% false discovery rate set where a peak was also reported (Li et al., 2008). We obtained 427 bound sequences in the Tcf1/Hnf1 data set (Odom et al., 2006) by implementing a program to perform the procedure described at http://jura.wi.mit.edu/young_public/hESregulation/Regions.html to the raw data. To create Figures 7A and 7B, we added 1 kb to either side of the ChIP-chip peak (for Caudal) or the center of the identified bound sequence (for Tcf1/Hnf1) and determined the relative enrichment in overlapping 500-base windows, using a 10-fold excess of 2 kb random genomic regions taken from the *Drosophila* genome (for Caudal) or the human genome (for Tcf1/Hnf1) as a background set.

Data Availability

Supplementary data, all original data files and array probe sequences are online at http://hugheslab.ccbr.utoronto.ca/supplementary-data/homeodomains1/ and http://the_brain.bwh.harvard.edu/pbms/webworks2/.

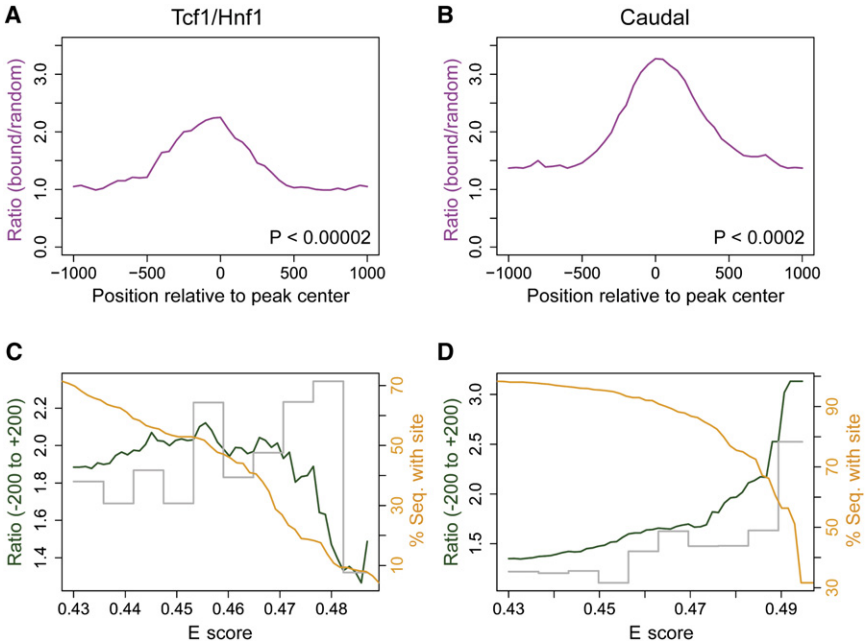

Figure 7. Enrichment of Sequences Preferred In Vitro within Genomic Sequences Bound In Vivo by the Same Protein

(A) Comparison of bound to randomly selected sequences for human Tcf1/Hnf1 (Odom et al., 2006), showing the relative enrichment of our 8-mers (at 0.456 cutoff). p value was calculated for the interval (−200 to +200) by the Wilcoxon-Mann-Whitney rank sum test, comparing the number of occurrences per sequence in the bound set versus the background set.

(B) Same as (A), but for *Drosophila* Caudal (Li et al., 2008) (at 0.493 cutoff).

(C) Relative enrichment (green line) in the −200 to +200 window as a function of the E score cutoff for Tcf1/Hnf1. The orange line shows the proportion of bound fragments with at least one such sequence in the same interval. The gray bars show the relative enrichment of 8-mers within each interval of 0.006, e.g., only 0.43–0.436 for the first interval.

(D) Same as (C), but for Caudal.

ACCESSION NUMBERS

Microarray data reported in this paper have been deposited in the NCBI GEO database with Platform ID GPL6760 and Series ID GSE11239.

SUPPLEMENTAL DATA

Supplemental Data include Supplemental Experimental Procedures, Results, and Discussion, nine figures, four tables, and Supplemental References and can be found with this article online at http://www.cell.com/cgi/content/full/133/7/1266/DC1/.

ACKNOWLEDGMENTS

This project was supported by funding from the Canadian Institutes of Health Research (MOP-77721), Genome Canada through the Ontario Genomics Institute, the Ontario Research Fund, and the Canadian Institute for Advanced Research to T.R.H., M.L.B. and G.B., and by grant R01 HG003985 from NIH/NHGRI to M.L.B. M.F.B. was supported in part by a U.S. National Science Foundation Graduate Research Fellowship. A.A.P. was supported in part by a National Defense Science and Engineering Graduate Fellowship from the U.S. Department of Defense and an Athinoula Martinos Fellowship from HST. S.A.J. was supported in part by a U.S. National Science Foundation fellowship Postdoctoral Research Fellowship in Biological Informatics. We thank Genita Metzler, Hanna Kuznetsov, Chi-Fong Wang, Anastasia Vedenko, Frédéric Bréard, David Coburn, Dimitri Terterov, Ally Yang, Harm van Bakel, Wing Chang, and John Calarco for technical assistance and Shoshana Wodak, Fritz Roth, Charlie Boone, Jack Greenblatt, Ben Blencowe, Bill Stanford, and Trevor Siggers for helpful discussions and critical evaluation of the manuscript.

Received: November 21, 2007
Revised: March 10, 2008
Accepted: May 12, 2008
Published: June 26, 2008

REFERENCES

Banerjee-Basu, S., Moreland, T., Hsu, B.J., Trout, K.L., and Baxevanis, A.D. (2003). The Homeodomain Resource: 2003 update. Nucleic Acids Res. *31*, 304–306.

Benos, P.V., Bulyk, M.L., and Stormo, G.D. (2002). Additivity in protein-DNA interactions:How good an approximation is it? Nucleic Acids Res. *30*, 4442–4451.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. *24*, 1429–1435.

Blackwell, T.K., Huang, J., Ma, A., Kretzner, L., Alt, F.W., Eisenman, R.N., and Weintraub, H. (1993). Binding of myc proteins to canonical and noncanonical DNA sequences. Mol. Cell. Biol. *13*, 5216–5224.

Bryne, J.C., Valen, E., Tang, M.H., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update. Nucleic Acids Res. *36*, D102–D106.

Bult, C.J., Blake, J.A., Richardson, J.E., Kadin, J.A., Eppig, J.T., Baldarelli, R.M., Barsanti, K., Baya, M., Beal, J.S., Boddy, W.J., et al. (2004). The Mouse Genome Database (MGD): Integrating biology with the genome. Nucleic Acids Res. *32*, D476–D481.

Carr, A., and Biggin, M.D. (1999). A comparison of in vivo and in vitro DNA-binding specificities suggests a new model for homeoprotein DNA binding in Drosophila embryos. EMBO J. *18*, 1598–1608.

Carroll, J.S., Liu, X.S., Brodsky, A.S., Li, W., Meyer, C.A., Szary, A.J., Eeckhoute, J., Shao, W., Hestermann, E.V., Geistlinger, T.R., et al. (2005). Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell *122*, 33–43.

Catron, K.M., Iler, N., and Abate, C. (1993). Nucleotides flanking a conserved TAAT core dictate the DNA binding specificity of three murine homeodomain proteins. Mol. Cell. Biol. *13*, 2354–2365.

Chan, S.K., and Mann, R.S. (1993). The segment identity functions of Ultrabithorax are contained within its homeo domain and carboxy-terminal sequences. Genes Dev. *7*, 796–811.

Chen, C.Y., and Schwartz, R.J. (1995). Identification of novel DNA binding targets and regulatory domains of a murine tinman homeodomain factor, nkx-2.5. J. Biol. Chem. *270*, 15628–15633.

Chen, X., Hughes, T.R., and Morris, Q. (2007). RankMotif++: A motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. Bioinformatics *23*, i72–i79.

Damante, G., Pellizzari, L., Esposito, G., Fogolari, F., Viglino, P., Fabbro, D., Tell, G., Formisano, S., and Di Lauro, R. (1996). A molecular code dictates sequence-specific DNA recognition by homeodomains. EMBO J. *15*, 4992–5000.

Dickinson, L.A., Joh, T., Kohwi, Y., and Kohwi-Shigematsu, T. (1992). A tissue-specific MAR/SAR DNA-binding protein with unusual binding site recognition. Cell *70*, 631–645.

Dudley, A.M., Aach, J., Steffen, M.A., and Church, G.M. (2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc. Natl. Acad. Sci. USA *99*, 7554–7559.

Ekker, S.C., von Kessler, D.P., and Beachy, P.A. (1992). Differential DNA sequence recognition is a determinant of specificity in homeotic gene action. EMBO J. *11*, 4059–4072.

Ekker, S.C., Jackson, D.G., von Kessler, D.P., Sun, B.I., Young, K.E., and Beachy, P.A. (1994). The degree of variation in DNA sequence recognition among four Drosophila homeotic proteins. EMBO J. *13*, 3551–3560.

Fraenkel, E., Rould, M.A., Chambers, K.A., and Pabo, C.O. (1998). Engrailed homeodomain-DNA complex at 2.2 A resolution: A detailed view of the interface and comparison with other engrailed structures. J. Mol. Biol. *284*, 351–361.

Furukubo-Tokunaga, K., Flister, S., and Gehring, W.J. (1993). Functional specificity of the Antennapedia homeodomain. Proc. Natl. Acad. Sci. USA *90*, 6360–6364.

Galant, R., Walsh, C.M., and Carroll, S.B. (2002). Hox repression of a target gene: Extradenticle-independent, additive action through multiple monomer binding sites. Development *129*, 3115–3126.

Han, Z., and Olson, E.N. (2005). Hand is a direct target of Tinman and GATA factors during Drosophila cardiogenesis and hematopoiesis. Development *132*, 3525–3536.

Hanes, S.D., and Brent, R. (1989). DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. Cell *57*, 1275–1283.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., et al. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99–104.

Jackson, M., Watt, A.J., Gautier, P., Gilchrist, D., Driehaus, J., Graham, G.J., Keebler, J., Prugnolle, F., Awadalla, P., and Forrester, L.M. (2006). A murine specific expansion of the Rhox cluster involved in embryonic stem cell biology is under natural selection. BMC Genomics *7*, 212.

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.

Kissinger, C.R., Liu, B.S., Martin-Blanco, E., Kornberg, T.B., and Pabo, C.O. (1990). Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: A framework for understanding homeodomain-DNA interactions. Cell *63*, 579–590.

Larroux, C., Fahey, B., Degnan, S.M., Adamski, M., Rokhsar, D.S., and Degnan, B.M. (2007). The NK homeobox gene cluster predates the origin of Hox genes. Curr. Biol. *17*, 706–710.

Laughon, A. (1991). DNA binding specificity of homeodomains. Biochemistry *30*, 11357–11367.

Li, M.Z., and Elledge, S.J. (2005). MAGIC, an in vivo genetic method for the rapid construction of recombinant DNA molecules. Nat. Genet. *37*, 311–319.

Li, X.Y., Macarthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C.L., et al. (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol. *6*, e27.

Liberzon, A., Ridner, G., and Walker, M.D. (2004). Role of intrinsic DNA binding specificity in defining target genes of the mammalian transcription factor PDX1. Nucleic Acids Res. *32*, 54–64.

Lin, L., and McGinnis, W. (1992). Mapping functional specificity in the Dfd and Ubx homeo domains. Genes Dev. *6*, 1071–1081.

Mann, R.S., and Chan, S.K. (1996). Extra specificity from extradenticle: the partnership between HOX and PBX/EXD homeodomain proteins. Trends Genet. *12*, 258–262.

Mao, C.A., Gan, L., and Klein, W.H. (1994). Multiple Otx binding sites required for expression of the Strongylocentrotus purpuratus Spec2a gene. Dev. Biol. *165*, 229–242.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., et al. (2003). TRANSFAC: Transcriptional regulation, from patterns to profiles. Nucleic Acids Res. *31*, 374–378.

Miller, J.C., and Pabo, C.O. (2001). Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. J. Mol. Biol. *313*, 309–315.

Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A., and Bulyk, M.L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat. Genet. *36*, 1331–1339.

Odom, D.T., Dowell, R.D., Jacobsen, E.S., Nekludova, L., Rolfe, P.A., Danford, T.W., Gifford, D.K., Fraenkel, E., Bell, G.I., and Young, R.A. (2006). Core transcriptional regulatory circuitry in human hepatocytes. Mol. Syst. Biol. *2*, 2006.0017.

Overdier, D.G., Porcella, A., and Costa, R.H. (1994). The DNA-binding specificity of the hepatocyte nuclear factor 3/forkhead domain is influenced by amino-acid residues adjacent to the recognition helix. Mol. Cell. Biol. *14*, 2755–2766.

Philippakis, A.A., Qureshi, A., Berger, M.F., and Bulyk, M.L. (2008). Design of compact, universal DNA microarrays for protein binding microarray experiments. J. Comput. Biol., in press.

Rajkovic, A., Yan, C., Yan, W., Klysik, M., and Matzuk, M.M. (2002). Obox, a family of homeobox genes preferentially expressed in germ cells. Genomics *79*, 711–717.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: An open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. *32*, D91–D94.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Svingen, T., and Tonissen, K.F. (2006). Hox transcription factors and their elusive mammalian gene targets. Heredity *97*, 88–96.

Treisman, J., Gonczy, P., Vashishtha, M., Harris, E., and Desplan, C. (1989). A single amino acid can determine the DNA binding specificity of homeodomain proteins. Cell *59*, 553–562.

Tucker-Kellogg, L., Rould, M.A., Chambers, K.A., Ades, S.E., Sauer, R.T., and Pabo, C.O. (1997). Engrailed (Gln50 → Lys) homeodomain-DNA complex at 1.9 A resolution: Structural basis for enhanced affinity and altered specificity. Structure *5*, 1047–1054.

Wilson, D.S., and Desplan, C. (1999). Structural basis of Hox specificity. Nat. Struct. Biol. *6*, 297–300.

Wolberger, C. (1996). Homeodomain interactions. Curr. Opin. Struct. Biol. *6*, 62–68.

Wolfe, S.A., Grant, R.A., Elrod-Erickson, M., and Pabo, C.O. (2001). Beyond the "recognition code": Structures of two Cys2His2 zinc finger/TATA box complexes. Structure *9*, 717–723.