Resource

Cell

DNA-Binding Specificities of Human Transcription Factors

Arttu Jolma,[1,2,8] Jian Yan,[1,8] Thomas Whitington,[1] Jarkko Toivonen,[3] Kazuhiro R. Nitta,[1] Pasi Rastas,[3] Ekaterina Morgunova,[1] Martin Enge,[1] Mikko Taipale,[2] Gonghong Wei,[2] Kimmo Palin,[2] Juan M. Vaquerizas,[4] Renaud Vincentelli,[5] Nicholas M. Luscombe,[4] Timothy R. Hughes,[6] Patrick Lemaire,[7] Esko Ukkonen,[3] Teemu Kivioja,[1,2,3] and Jussi Taipale[1,2,*]

[1]Science for Life Center, Department of Biosciences and Nutrition, Karolinska Institutet, 141 83 Huddinge, Sweden
[2]Genome-Scale Biology Program
[3]Department of Computer Science
University of Helsinki, 00014 Helsinki, Finland
[4]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK
[5]Architecture et Fonction des Macromolécules Biologiques, UMR7257 CNRS, Université Aix-Marseille, 163 Avenue de Luminy, 13288 Marseille Cedex 9, France
[6]Donnelly Center, Banting and Best Department of Medical Research and Department of Molecular Genetics, University of Toronto, Ontario M5S 3E1, Canada
[7]CRBM, 1919 Route de Mende, 34293 Montpellier, France
[8]These authors contributed equally to this work
*Correspondence: jussi.taipale@ki.se
http://dx.doi.org/10.1016/j.cell.2012.12.009

SUMMARY

Although the proteins that read the gene regulatory code, transcription factors (TFs), have been largely identified, it is not well known which sequences TFs can recognize. We have analyzed the sequence-specific binding of human TFs using high-throughput SELEX and ChIP sequencing. A total of 830 binding profiles were obtained, describing 239 distinctly different binding specificities. The models represent the majority of human TFs, approximately doubling the coverage compared to existing systematic studies. Our results reveal additional specificity determinants for a large number of factors for which a partial specificity was known, including a commonly observed A- or T-rich stretch that flanks the core motifs. Global analysis of the data revealed that homodimer orientation and spacing preferences, and base-stacking interactions, have a larger role in TF-DNA binding than previously appreciated. We further describe a binding model incorporating these features that is required to understand binding of TFs to DNA.

INTRODUCTION

Understanding of transcriptional networks that control animal development as well as physiological and pathological processes requires the cataloging of target genes of each transcription factor (TF) under all possible developmental and environmental conditions. Approaches identifying central TFs and their target genes in simple models where environmental conditions are stable, such as early embryonic development of sea urchin, *C. elegans*, and *Drosophila*, have been successful (Davidson and Levine, 2008; Walhout, 2011). Similar approaches can also be applied to analysis of human transcriptional networks important for particular processes, using methods such as classical genetics, chromatin immunoprecipitation followed by sequencing (ChIP-seq), and RNAi (see, for example, Balaskas et al., 2012; Chen et al., 2008; Chia et al., 2010). However, due to the large number of TFs (>1,000; Vaquerizas et al., 2009), cell types, and environmental states, exhaustive application of such approaches to understand human transcriptional regulation is not feasible.

Furthermore, observing where TFs bind in the genome does not explain why they bind there. To understand TF binding, it is necessary to develop a model that is based on biochemical principles of affinity and mass action (e.g., Hallikas et al., 2006; Segal et al., 2008). Such a model would allow reading of the regulatory genetic code, and prediction of gene expression based on sequence. It would also be very important for personalized medicine because it would allow prediction of the effects of previously unknown variants or mutations on gene expression and disease susceptibility (Tuupanen et al., 2009). The parameters of such a model include the initial concentrations and the quantitative binding specificities of DNA-binding proteins such as histones (Kaplan et al., 2009) and all TFs encoded by the human genome.

A binding specificity model for a TF should describe its affinity toward all possible DNA sequences. By assuming that each TF-DNA base interaction is independent (Benos et al., 2002; Roulet et al., 2002), TF-binding specificity can be expressed as a position weight matrix (PWM), which describes the effect of each base on binding separately. Due to the low resolution of most existing data (Jolma and Taipale, 2011), it is not clear how generally applicable this model is (Badis et al., 2009; Zhao and Stormo, 2011).

Despite the central importance of transcriptional regulation in development and disease, very little work has concentrated on

analysis of binding specificities of human TFs. Previous systematic studies have concentrated on specificities of TFs from common model organisms, including yeast, *C. elegans*, *Drosophila*, and mouse (Badis et al., 2009; Berger et al., 2008; Grove et al., 2009; Noyes et al., 2008). In general, they have analyzed bacterially expressed TF-DNA-binding domains (DBDs), with very few studies analyzing full-length TFs.

In this work, we have systematically analyzed specificities of most human TFs using a high-throughput SELEX (HT-SELEX) (Jolma et al., 2010; Jolma and Taipale, 2011; Oliphant et al., 1989; Tuerk and Gold, 1990). Comparison of 79 pairs of experiments for full-length TFs and their DBDs revealed that in general, the DBD defines the primary DNA-binding specificity. Analysis of the data revealed that the vast majority of interactions that occur between a TF and individual DNA bases are independent of each other. However, strong base interdependencies were observed in a stretch of three to five A or T residues flanking the core binding site in multiple TF classes, consistent with proposed shape-based recognition of DNA (Rohs et al., 2010). Adjacent bases also deviated more from the independent model than bases that were farther apart, indicating that base-stacking interactions have a larger role in TF-DNA binding than what has been previously appreciated. We also commonly observed formation of dimers, with strong orientation and spacing preferences. These preferences could be used to further classify TF subfamilies that had identical primary specificities. We show that models incorporating adjacent dinucleotides and dimer spacing and orientation preferences improve modeling of TF binding to DNA and that the dimer model can be generalized to analyze large heteromeric TF-DNA complexes.

RESULTS

Genome-Scale TF-DNA-Binding Specificity Assay

To determine the binding specificities of mammalian TFs, we cloned 891 human and 444 mouse DBDs and 984 human full-length TFs into Gateway recombination vectors and expressed the corresponding C-terminally tagged proteins in mammalian cells. As a control, a subset of these proteins was also expressed in *E. coli* as N-terminal fusions (see Table S1 available online).

The sequences that the TFs bind to were then determined by HT-SELEX (Figure 1A). Robust enrichment of specific sequences was observed for 303 human DBDs, 84 mouse DBDs, and 151 human full-length TFs, representing 411 different TFs (Table S1). In general, a high fraction of experiments was successful for most TF families (Table S2). Of the large TF families comprising more than 30 factors, two had a low success rate: high-mobility group (HMG), and C2H2 zinc finger proteins. The results are consistent with many HMG proteins not binding DNA sequence specifically (Stros et al., 2007) and with earlier observations that many C2H2 zinc finger proteins do not bind specific DNA sequences in protein-binding microarray (PBM) experiments (T.R.H., unpublished data). C2H2 domains are also known to be used for other purposes than DNA binding, even in proteins that also contain DNA-binding C2H2 zinc fingers (Brayer and Segal, 2008; Brown, 2005).

To determine primary binding specificities for the factors, we built a PWM from enriched subsequences using a multinomial method we have described previously (Jolma et al., 2010; Figure 1B; Table S3). Matrices were corrected for nonspecific DNA carryover. The matrices generated using this method from early SELEX cycles were generally similar to those obtained by a ratio method, where normalized subsequence counts observed in a given cycle are divided by normalized counts observed in the previous cycle (Figure 1C).

We have previously established that many TFs that bind DNA as monomers can also bind as homodimers and that the dimers display strong orientation and spacing preferences (Jolma et al., 2010). To analyze homodimeric binding globally, we analyzed the enriched sequences to identify TFs that bound to two similar sites within a single DNA fragment. The cases where the dimers displayed clear orientation and spacing preferences were included in the set of PWMs analyzed further. In total, we obtained 830 binding profiles for human and mouse TFs (Table S3).

Full-Length TFs and Isolated DBDs Bind Similar Sequences

We next analyzed the similarity between the obtained binding specificities for full-length proteins and the corresponding profiles for DBDs using the minimal Kullback-Leibler divergence (KL) method (Wei et al., 2010). Analysis of profiles for all the 79 human TFs for which both full-length and DBD experiments were successful revealed that in the vast majority of cases, the full-length and DBD PWMs were very similar (KL <2). Most differences between the models were minor (Figure 1D), being generally of similar magnitude than those observed between replicate experiments (KL, 0.51 ± 0.32). The only clear difference identified affected a homodimeric site for the ETS factor ELK1 (Figure 1D). These results suggest that in most cases, analysis of DBDs is sufficient for determination of TF-binding specificities.

Analysis of Model Width and Coverage

Analysis of the length and information content of the PWMs revealed that on average, they were 13 bp long and contained 15.6 bits of information (Figure 2A; data not shown). There was a clear correlation between width of the binding profile and its information content (data not shown), and clear decrease of information content per base was not observed in longer motifs.

We next determined the fraction of high-confidence human TFs that are covered by models in our data and in existing databases, including a literature-curated set (JASPAR; Portales-Casamar et al., 2010), and a collection based on a high-throughput approach (PBMs; Badis et al., 2009; Berger et al., 2008). This analysis revealed that our data covered approximately two times larger number of human TFs than PBMs, the largest currently available systematic data set (Figure 2B). Because PBM analyses have focused mostly on mouse TFs, we also compared coverage based on protein similarity, again revealing that our data set is clearly the largest collection of mouse or human TFs, covering more than 50% of all high-confidence TFs at a 90% similarity threshold (Figure 2B).

To analyze the differences between the PBM and SELEX data in more detail, we compared separately the number of TFs (mouse and human) belonging to different structural TF families. For eight TF families that primarily bind DNA as monomers,

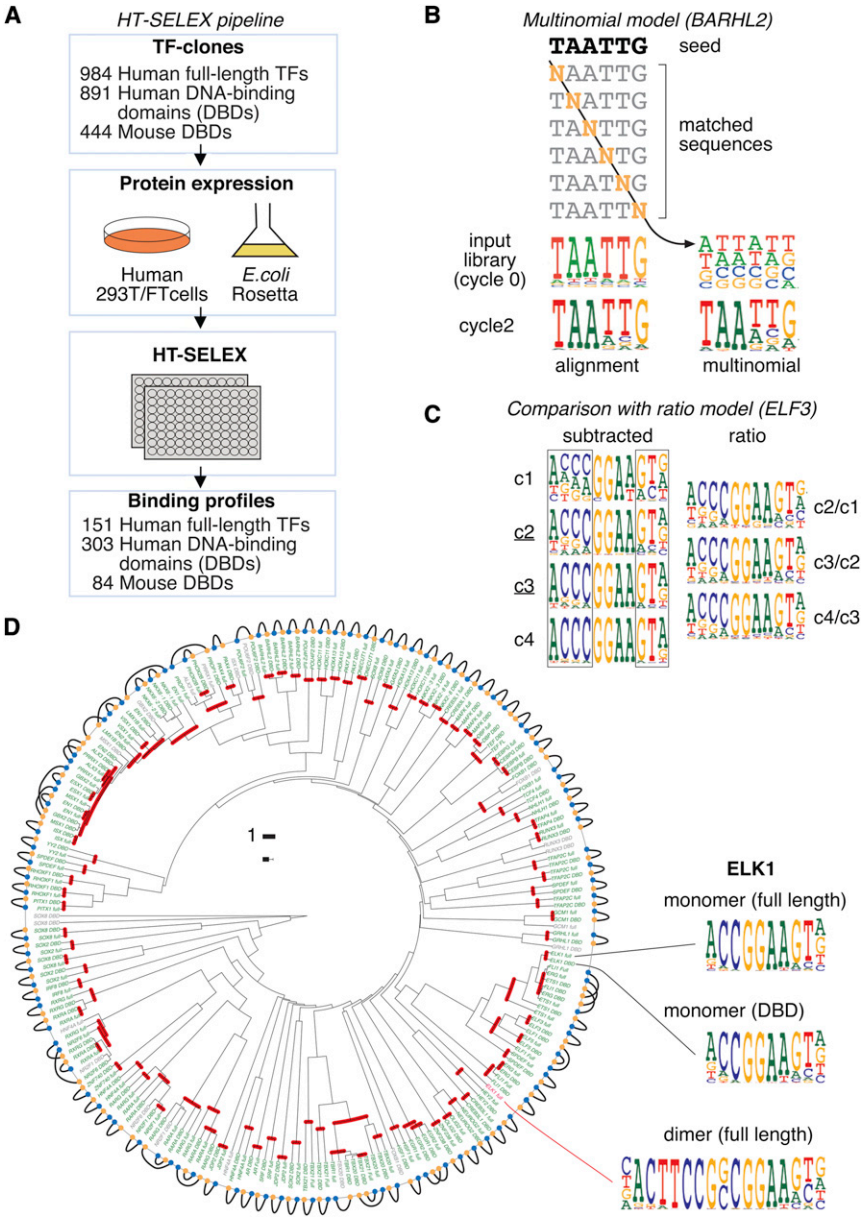

Figure 1. Analysis of TF-Binding Specificity

(A) Genome-Scale HT-SELEX pipeline.

(B) PWM generation using the multinomial algorithm. Multinomial model is generated by counting the occurrences of each base at a given position when all other bases exactly match a seed sequence. Note that simple alignment generates an excessively stringent model, resembling the consensus even when random sequences (input library) are analyzed.

(C) Comparison between binding profiles for ELF3 DBD obtained using background subtraction (left) and count ratio (right) methods. Note that models generated using background subtraction are too loose at cycle 1 (c1) due to saturation of high-affinity sites and that by cycle 4 (c4), they become excessively stringent due to exponential enrichment. However, at cycles 2 and 3 (underlined), the background subtraction model is similar to a ratio model (right, the cycles between which the ratio was calculated are also indicated). Note that the choice of SELEX cycle has the largest effect on bases that have moderate effect on binding (boxes).

(D) Binding profiles obtained using full-length proteins are very similar to those obtained using the corresponding DBDs. Bars indicate divergence of 1 and divergence between PWMs from replicate experiments ±SD. Dendrogram shows all PWM models for the same protein in DBD (orange) and full-length (blue) form. Black arcs connect a DBD model to its corresponding full-length model, and red lines indicate the dendrogram branchpoint. Some secondary PWM models (gray) were generated only for a DBD or full-length protein due to weaker enrichment in the other sample. Logos highlight the only clear difference found between DBD and full-length models.

See also Figure S4 and Table S1.

a similar number of models were described (Figure 2C). However, for the remaining 23 families that bind DNA mostly as dimers or multimers in HT-SELEX, dramatically higher number of models were obtained (Figure 2D). These differences appear to be related to the fact that PBMs contain all 10 bp sequences, whereas 14–40 bp random sequences are used in HT-SELEX. This results in either failure of PBM analyses to identify long binding sequences or recovery of a partial specificity or a half-site of a dimer (Figures 2E and 2F; Figure S1).

Different Structural Families of TFs Have Clearly Distinct Specificities

We next generated a network where TFs were connected to each other if their HT-SELEX PWM models were similar (Figure 3; see Experimental Procedures). In this analysis, the different TF structural families separated into distinct subnetworks (Figure 3; for larger images and logos, see Data S1). Only three exceptions were found: GMEB2, SNAI2, and CPEB1. In each case, a single factor from one structural family associated with a group of factors from another family (Figure 3).

Because many of the PWMs were similar, we used a minimum dominating set of the network to identify 239 PWMs that could describe the entire set of profiles. Several large groups of TFs that could be represented by a single PWM were identified, including ETS class I proteins, and subsets of homeodomain and bHLH proteins that bound to canonical TAATTA and CACGTG sites, respectively (Figure 3). The obtained PWMs for the entire set of 146 homeodomain, 39 bHLH, and 24 ETS proteins could be described by only 53, 9, and 10 representative models, respectively. In contrast, 42 distinct profiles were required to describe 53 C2H2 zinc finger proteins. Many of the zinc finger models, including those for Zfp652,

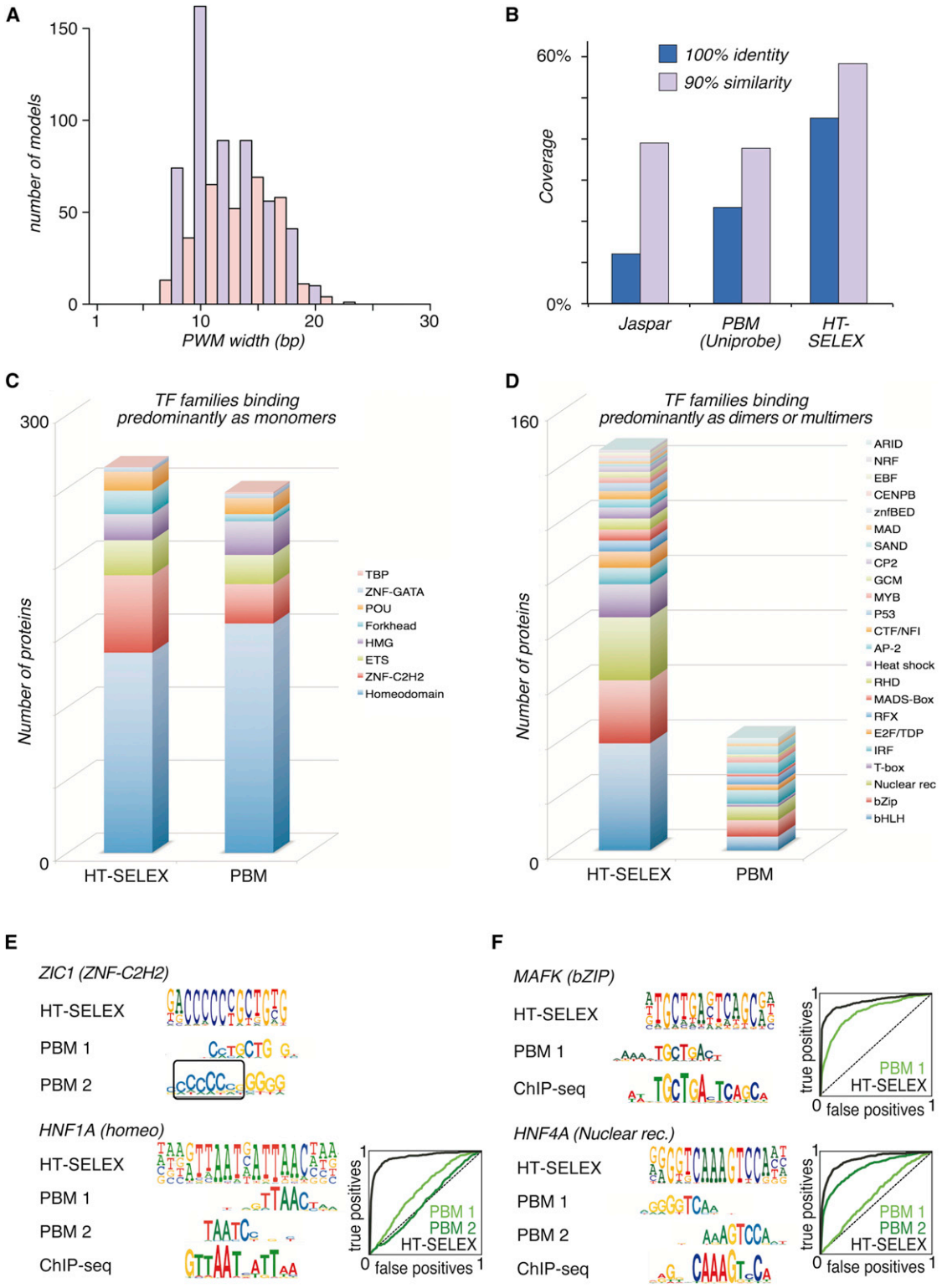

Figure 2. Comparison of Coverage of TFs

(A) Histogram showing the distribution of PWM model widths. Note that TFs prefer even (blue) over odd (red) widths due to palindromic sites and that a width of 10 bp corresponding to a single turn of a DNA helix is the most common. Note also that the specificity of most TFs extends beyond 10 bp.

(B) Coverage of human high-confidence TFs by JASPAR CORE (left bars), PBM (middle bars), and HT-SELEX (right bars) at indicated thresholds.

(C) Number of TFs for which a model has been derived using PBM or HT-SELEX. Colors indicate different structural TF families that bind DNA primarily as monomers.

(legend continued on next page)

ZNF410/APA1, ZKSCAN3/ZNF306, ZNF282, ZNF232, ZBTB49/ZNF509, ZNF524, and ZNF713, that we identify here were dissimilar to any model described previously (Data S1).

We also generated a similar network that included also existing literature curated and PBM data on human and mouse TF specificities (Figure S2). This analysis revealed that our data were in broad agreement with the more limited information on TF-binding specificities that had been described before. Most clear differences could be explained by the lower resolution of the previously used methods (Figure S1A), shorter sequence length analyzed (Figure S1B), or issues related to the conversion of raw PBM data into PWM form (Figure S1C).

Conservation of Binding Specificities

Analysis of HT-SELEX-derived PWMs revealed that in all tested cases, the mouse and human ortholog-binding specificities were similar (Figure 3, compare triangles and circles). The lack of differences was not due to our inability to detect them because we did identify a difference that was caused by a missense mutation in our Egr1 clone. The mutation affects a DNA-contacting residue but is not found in mouse reference genome or SNPs, indicating that the mutation is either private or introduced in cloning (Data S1; Table S1).

Classification of TFs Based on Their DNA-Binding Specificities

We have previously classified the ETS family of TFs into four classes based on two independent analyses of their binding preferences (Wei et al., 2010). Our SELEX analysis of 24 members of the 27 ETS family TFs corroborates these four classes (Figure 4A). However, even within this well-studied group of factors, we could identify additional novel specificity determinants for three out of the four classes (Figure 4A; Data S1).

We could also identify other families that displayed clear one-to-one relationships between proteins and binding specificity models. For example, five classes of GLI-like C2H2 zinc fingers, four main classes of basic-helix-loop-helix (bHLH) proteins, four classes of PAX proteins, and two classes each of E2Fs, HSFs, MADS proteins, CUT+homeodomains, and SP/KLF/EGR C2H2 zinc fingers could be clearly identified (Figure 3; Data S1).

Classification of TFs Based on Dimer Spacing and Orientation

Dimer orientation and spacing preferences could be used to further classify some factors that showed similar monomer-binding specificities. For example, the ETS class I factors ERG, ETS1, and ELK1 preferred to bind to different homodimeric sites (Figure 4A; see also Babayeva et al., 2010; Jolma et al., 2010; Lamber et al., 2008). Similarly, both T box factors and forkhead proteins displayed one type of monomer specificity but seven and three distinct dimeric spacing/orientation preferences, respectively (Figure 4B; Data S1).

In some cases, both spacing and orientation preferences, and the monomer sites/half-sites, could vary. For example, RHD family factors could be classified to NFAT and NF-κB subgroups based on half-site specificity, and the NFAT subgroup further diverged to two distinct orientation and spacing preference groups (Data S1). Similarly, nuclear receptors could be classified to 12 groups, based on eight different half-site specificities and five different spacing groups within factors that specifically bound one type of half-site (Data S1). Homeodomains could also be subclassified based on monomer specificity and spacing and orientation preferences (Figure 3; Data S1).

For posterior homeodomains (Data S1), POU+homeodomains (Data S1), and bZIP proteins (Figure 4C; Data S1), classification was more complex because factors shared partially overlapping sites. For example, many bZIP proteins could bind to two distinct sites and be classified based on the sets of sites that they bind to. Their specificities were arranged in a tiled pattern, based on both overlapping half-site and spacing preferences (Figure 4C).

Independence of DNA Base Positions in TF Binding

To analyze how independently different base pairs bind to TFs, we compared observed counts of nucleotide pairs to the corresponding nucleotide pair counts expected based on a PWM (Figure 5A). Plotting of the observed counts against the expected counts revealed that the PWM was a good model for the vast majority of position pairs (Figure 5B).

Furthermore, calculation of the correlation between the nucleotide pair counts observed and predicted from the PWM for each pair of bases in all TF models revealed that only 0.9% of all pairs had a correlation coefficient that was lower than 0.9 (data not shown). PWM was particularly effective at modeling bases separated by more than three bases. Bases that were closer together displayed a somewhat larger deviation from the PWM model, with the largest difference observed for directly adjacent bases, with 5% of counts deviating from expected by more than 2-fold (Figure 5C; data not shown). These results indicate that TFs in general bind to base pairs independently of each other and that the strongest deviations from this model affect adjacent bases.

Deviations from the PWM Model

Although the PWM model explained pairs of bases well in most cases, some pairs displayed more than 5-fold deviations (expected/observed) from the PWM-based predictions. Such pairs were identified in several structural TF families.

The most striking case was SOX proteins. All SOX proteins bound to head-to-head pseudopalindromic sites (Data S1),

(D) Number of TFs for which a model has been derived using PBM or HT-SELEX. Colors indicate different structural TF families that bind DNA primarily as dimers or multimers in HT-SELEX.

(E) PBM identifies only partial specificities for TFs with long binding sites. HT-SELEX, PBM primary (PBM 1), PBM secondary (PBM 2), and ChIP-seq models are shown. Box indicates sequence that is misaligned to generate a palindromic PBM site that is inconsistent with SELEX.

(F) PBM identifies only half-sites for TFs that bind DNA as homodimers.

Insets in (E) and (F) are ROC curves showing enrichment of specific ChIP-seq peaks by the different in vitro PWMs.

See also Figure S1 and Table S2.

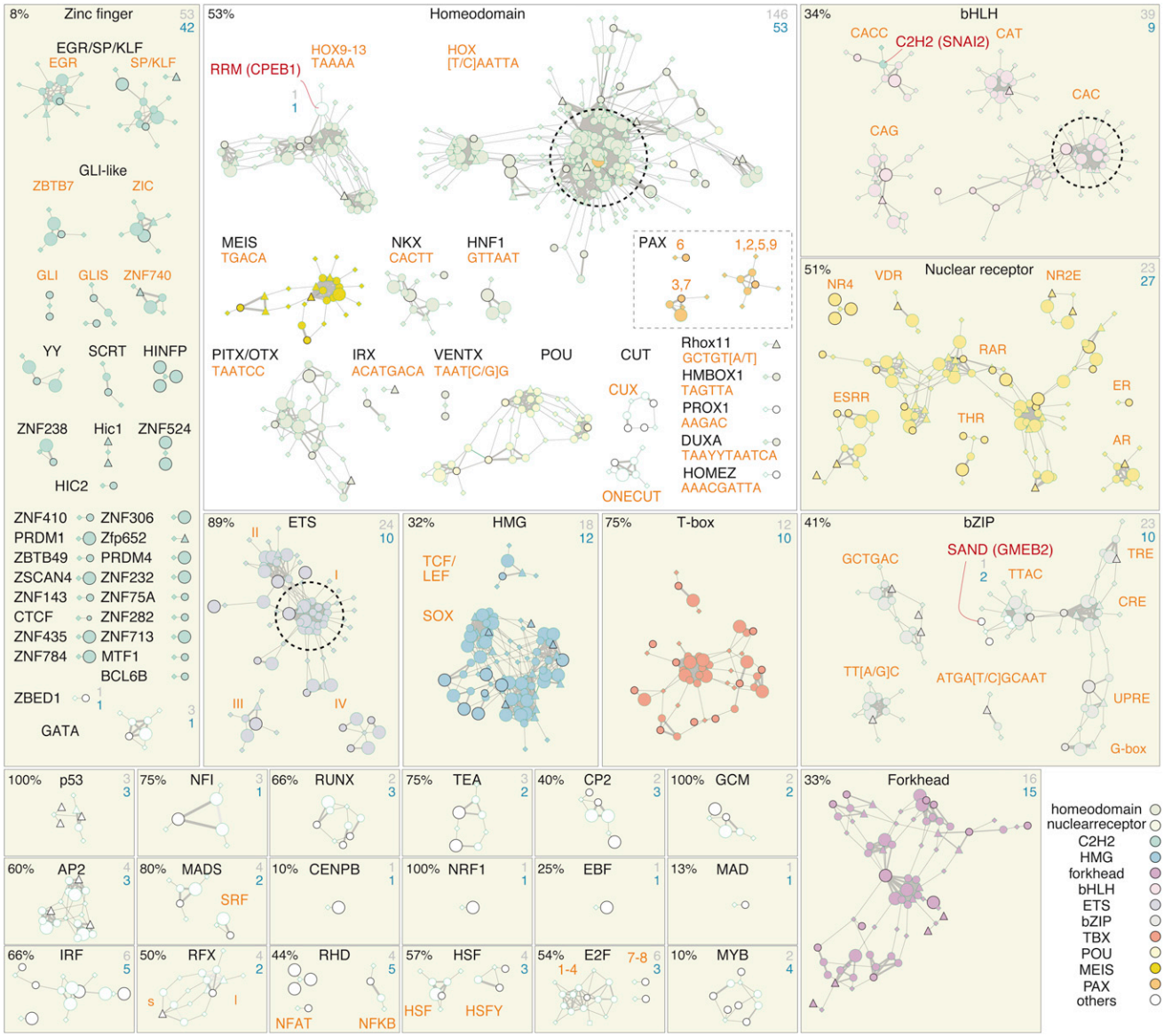

Figure 3. Network Representation of the Similarity of the Obtained PWMs

Diamonds indicate TF genes, and other nodes indicate individual PWMs; colors indicate TF family (bottom right). Models for human full-length TFs (large circles), DBDs (small circles), and mouse DBDs (triangles) are shown; representative models are indicated by black outline. Edges are drawn between a TF and its models, and between similar models. Subnetworks are named by family; where necessary, subfamilies are indicated with numbers or partial consensus sequences (orange typeface). Note that TFs cluster almost exclusively with other TFs of the same family (boxes; box in dotted line indicates that only some PAX proteins contain homeodomain). The three cases where a member of a class is included in a subnetwork composed of members of another class are indicated by red typeface. Fraction of TFs with models (top left of each box), total number of models (top right, above), and number of representative models (below) are also shown for each family. The three largest groups of models that are very similar to each other are circled (dotted line). See also Figure S2, Table S3, and Data S1.

which displayed an extremely strong correlation (>100-fold difference) between a dinucleotide that was present in one half-site with the corresponding dinucleotide in the other half-site, even though they were 9 or 10 bp apart. This effect is probably not mediated by a protein dimer but by base pairing in a stem loop formed from single-stranded DNA (Figure S1D).

We could further identify four different sources of correlations between bases. The first two types were associated with dimeric binding. The first was characterized by asymmetric binding of monomers in a tightly packed dimer (e.g., FLI1, MEIS2, PKNOX2) and could be modeled with a PWM that is nonpalindromic (Figure S3). The second type was due to the ability of some factors to bind to two distinct half-sites (e.g., HNF4A, many bZIP factors; data not shown).

The third type of base pair interdependency was linked to DNA binding by the homeodomain recognition helix. Strong

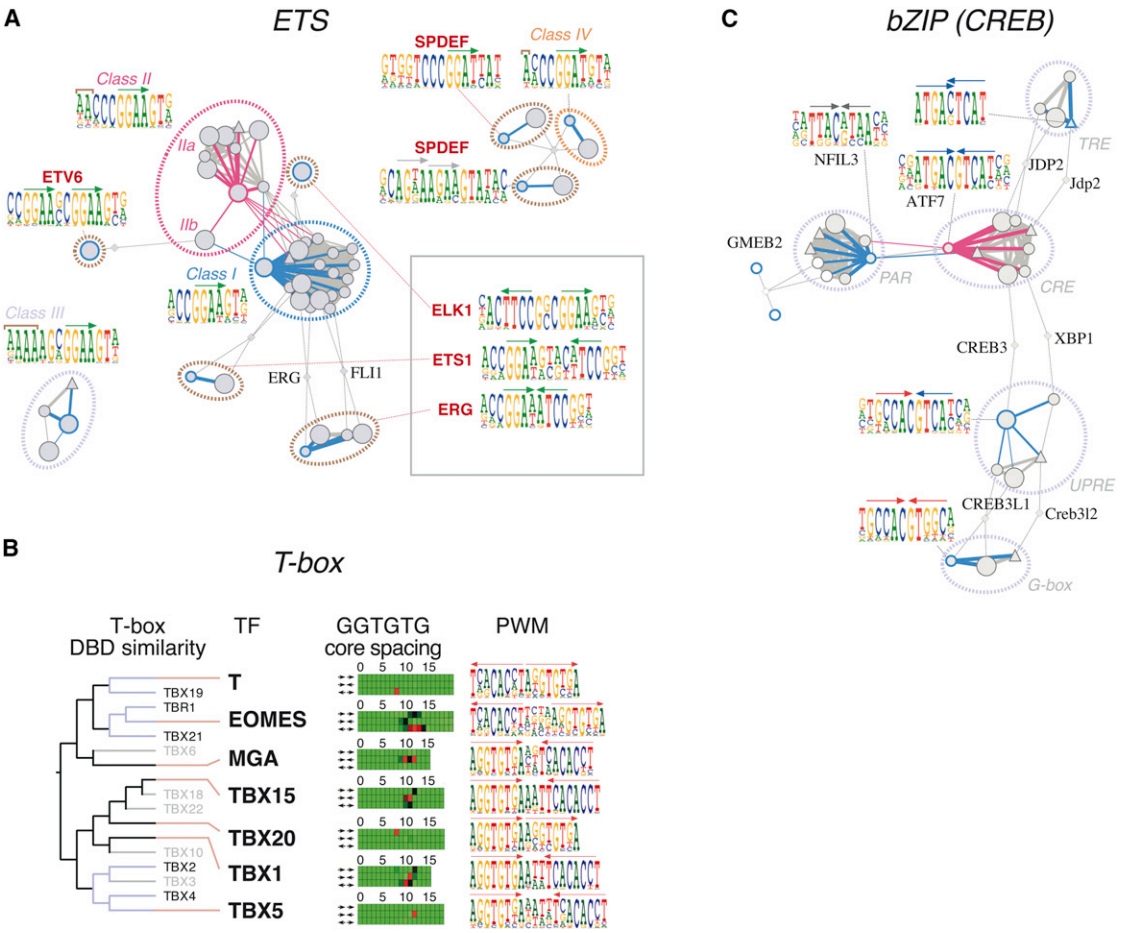

Figure 4. Classification of TFs Based on Their Binding Profiles

(A) ETS factors. Network analysis similar to that shown in Figure 3 indicates that HT-SELEX can accurately identify the four known ETS subclasses (indicated by colored ovals). Additional specificity determinants in classes II, III, and IV are indicated by brown brackets, and a novel dimer in ETV6 (class II) and two novel putative dimers in SPDEF (class IV) are indicated by brown dotted lines. Box indicates three different homodimeric sites within class I. Logos for representative PWM models are shown; green and gray arrows indicate GGA(A/T) and AGAA sequences, respectively.

(B) Classification of T box TFs based on dimer orientation and spacing. Left panel shows amino acid similarity dendrogram of T box DBDs. TFs for which models were not obtained are in gray. Middle panel shows heatmap displaying spacing and orientation (arrows) preferences of the enriched GGTGTG subsequences (red indicates max counts; green indicates 0); scale represents distances between the subsequence starting points. Right panel shows PWM describing most enriched dimeric binding site for each TF.

(C) A subset of bZIP TFs recognizes two types of target sites in a tiled pattern, covering four site types. Arrows above the logos indicate half-sites; black specifies TTAC, blue designates ATGAC, and red shows GCCAC. Note that JDP2, CREB3, XBP1, CREB3L1, and Creb3l2 each can bind to two different site types, forming a tiled pattern ranging from TRE element (top) to G box. Most TF nodes in (A) and (C) are omitted for clarity; for details, see Data S1.

correlations between adjacent bases were observed for BARHL2 (Figure 5A). Similarly, all posterior homeodomains (HOX9–HOX13) displayed strong correlations between bases located 5′ of the shared TAAAA subsequence (Figure 6A).

The fourth type of binding poorly explained by a PWM was the flanking of many TF core sequences with a stretch of three to five A or T bases (Figure 6B). Such sequences are predicted to narrow the minor groove of DNA, a feature that has been linked to shape-based DNA recognition (Rohs et al., 2010). Consistently, sequences favoring a narrow minor groove such as TTT or AAA were enriched much more than combinations of the same bases that result in much wider minor groove (Figure 6C; data not shown). Such A or T stretches also affected TF-DNA binding in vivo; core sequences enriched in ChIP-seq peaks for SPI1 (Wei et al., 2010), MAFG, and E2F7 (Figure 6B) were flanked with multiple As.

Models that Take into Account Deviations from the PWM Model

Given that adjacent nucleotides can affect each others' binding to a TF, and that many TFs bind to sequences that cannot be modeled by a standard mononucleotide model (PWM, a zero-order Markov model), we next tested whether the A stretch sequences could be explained by a model that takes into account adjacent bases. We first generated an adjacent dinucleotide model (ADM) for E2F3 from dinucleotide pair data. The

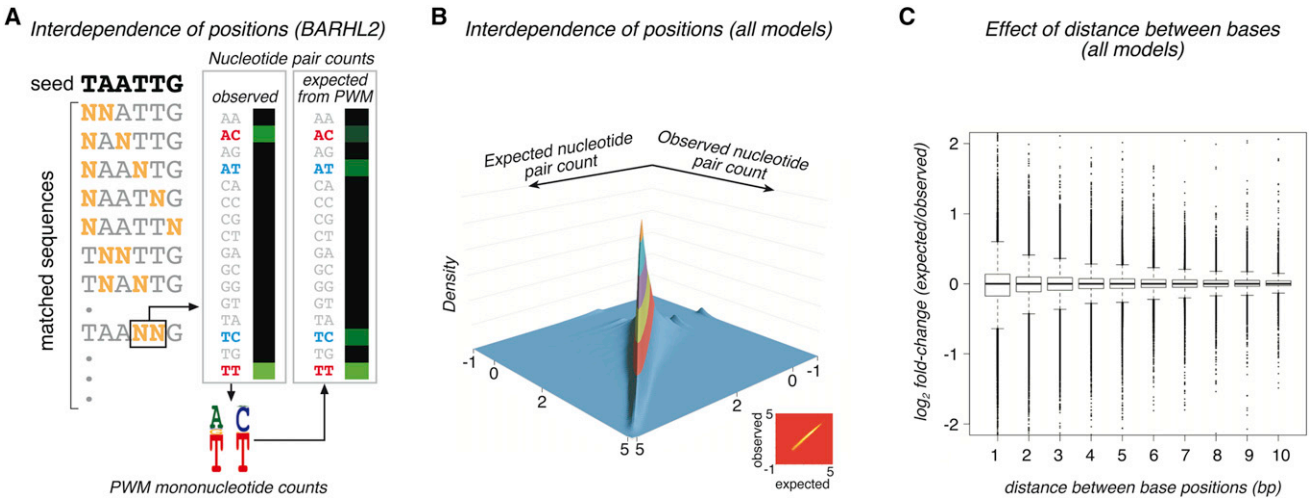

Figure 5. Global Analysis of Base Interdependency

(A) Analysis of interdependence of base positions. Nucleotide pair counts were generated for each pair of bases in such a way that bases that were not counted exactly matched the seed (left). Observed counts for each pair were then compared to those expected from mononucleotide distribution (bottom). Note that mononucleotide distribution cannot be used to generate accurate nucleotide pair counts for BARHL2-binding positions 4 and 5 (heatmaps; black is low, and green is high) due to a preferential binding of BARHL2 to taaACg or taaTTg (red) over taaATg and taaTCg (blue).

(B) In general, bases bind to TFs independently of each other. A density plot of counts observed versus counts expected from a PWM model for all possible pairs of base positions within all of the models generated in this study. Density (z axis; indicated both by height and by colors for clarity) of points in the x-y plane ($\log_{10}$ counts) is extremely concentrated at the diagonal, indicating that the vast majority of positions do not materially affect binding at other positions. Inset shows heatmap of the same data.

(C) A boxplot showing $\log_2$ fold change of count expected from a PWM model over observed count as a function of distance of the analyzed bases indicates that adjacent bases have stronger effect on each other than bases that are farther apart. Boxes indicate the middle quartiles, separated by median line. Whiskers indicate last values within 1.5 times the interquartile range from the box.

ADM is a series of first-order Markov chains that allows scoring of k-mers that are shorter than the model itself (Table S4). Plotting of the observed 10-mer counts for E2F3 against those expected from both PWM and ADMs revealed that the ADM was better at modeling the enrichment of 10-mer subsequences than a standard PWM (Figures 7A and 7B).

We next tested whether orientation and spacing preference matrix could be used to improve prediction of sequences enriched by TBX20, a factor that binds to a dimeric site where the same monomer is found in multiple different orientation and spacing configurations. For this purpose, we generated expected-observed plots for all possible combinations of two 4-mers with gaps of different length between them (gapped 8-mers). A model that incorporated spacing and orientation preferences (Table S4) described enriched gapped 8-mers much better (R^2 = 0.67 compared to 0.44) than a simple PWM (Figures 7C and 7D).

DISCUSSION

We report here high-resolution DNA-binding specificity for a large fraction of human TFs. Given the fact that proteins related in amino acid sequence generally bind to similar sites, we estimate that this resource represents the majority of all human TF-binding specificities. We also identify additional determinants of specificity for many factors for which a partial binding specificity was known before. The models described here are generated from a large number of sequences (average >7,000) and are of higher resolution than the existing SELEX-derived PWM models, which are affected by much higher Poisson error due to the low number of sequences analyzed (mostly 10–50).

Prior to this work, very few experiments have addressed binding specificities of human full-length TFs. Out of the 151 human full-length TFs that we obtained profiles for, previous high-resolution binding data exist only for ETS1 and GABPA (Wei et al., 2010). Of the 303 human DBDs we model here, 22 have been profiled previously (Portales-Casamar et al., 2010). Previous data for 78 and 311 TFs exist from human or mouse, respectively (Badis et al., 2009; Berger et al., 2008; Wei et al., 2010). Of all the 830 PWMs, 406 are similar to 1 or more of the 500 PWMs described before for homologous TFs; the remaining 424 profiles, representing 228 TFs, were different from any model that has been described before (Figure S2; SSTAT covariance $<1.5 \times 10^{-5}$).

Much of the existing data are derived using PBMs containing all possible 10 bp subsequences (Berger et al., 2006). Our results are generally in good agreement with the PBM data for TFs that bind to short sites. However, we find here that more than half of all binding models for TFs are >10 bp in length, suggesting that specificity of many TFs cannot be fully determined using PBMs. Consistently with this, the coverage of PBM models is very low for families that bind to DNA as dimers, and in many cases, the reported PBM model describes partial specificity or half-site. Many dimeric sites identified by HT-SELEX in this work had been identified before and/or were

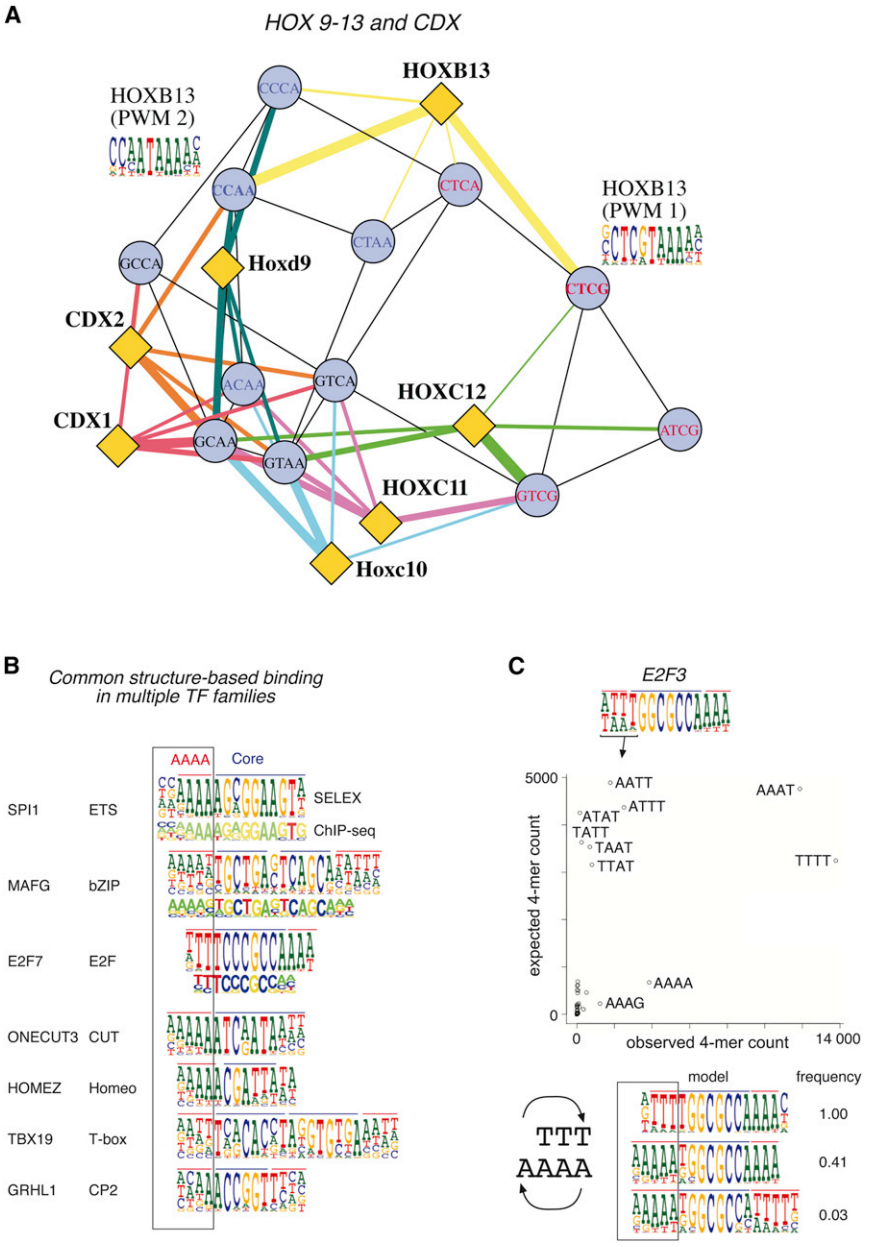

Figure 6. Examples of Base Pair Interdependencies in TF-DNA Binding

(A) Posterior homeodomains exhibit strong correlations between bound positions. Diamonds represent the indicated posterior homeodomain proteins, and circles represent enriched 9-mer sequences (circles, first four bases shown, last five bases are TAAAA). Edges are drawn between k-mer nodes if their Hamming distance is 1, and between a protein and a k-mer node if the k-mer is enriched by the protein. Edges between protein and k-mer nodes are colored for clarity, and their thickness represents the extent of the enrichment. Logos indicate two different PWM models for HOXB13 that are built using nonoverlapping sequences (blue and red).

(B) A stretch of A or T bases (box, red line above logos) is commonly observed adjacent to core TF-binding sites (blue line). Models generated using ChIP-seq (short) followed by motif discovery are shown below HT-SELEX-generated models (tall). SPI1 motif is from Wei et al. (2010).

(C) The bases are not independently bound but, instead, display a preference for a stretch of either A or T. Expected-observed plot for E2F3 describing 4-mers that precede the sequence GGCGCC. Note that AAA(T) and TTT(T) are strongly preferred over combinations such as AAT(T). The (T) is part of the E2F3 core. Bottom view shows binding motifs (middle) representing the three enriched combinations of core and flanking sequences and their relative frequencies (right).

See also Figure S3.

validated by ChIP-seq (Figure S4), indicating that HT-SELEX allows analysis of multimeric binding sites spanning 20 bp or more, which is beyond the capacity of any unbiased array technology.

TF-DNA-Binding Specificity Is Determined by the DBD

Some previous studies analyzing individual proteins have found that a TF and its isolated DBD bind to similar sequences (see, for example, Badis et al., 2009; Wei et al., 2010). On the other hand, some reports have found differences even between splicing variants of the same TF (Giguère et al., 1995). Most in vitro analyses of TF binding to date have analyzed specificity of isolated DBDs, whereas in vivo methods such as ChIP-seq have necessarily studied full-length TFs. Based on these results, it has not been possible to determine whether full-length TFs bind to different sites than isolated DBDs because observed differences in binding could be due to the differences in protein length or experimental method. Our analysis of 79 pairs of full-length TFs and DBDs revealed that the primary binding specificity of TFs is defined by the DBD. We found only one case, ELK1, where the specificity of full-length TF and DBD was clearly different.

Conservation of Binding Specificities

TF-binding specificities evolve very slowly (see, for example, Amoutzias et al., 2007; Bohmann et al., 1987; Struhl, 1987). Nevertheless, some examples of divergence of specificity exist in the literature (Solano et al., 1995), and systematic analysis of the divergence of specificities using current data has been hampered by the fact that the observed differences could be due to the different methods used to study orthologous TFs. Despite the morphological differences between mouse and human, we did not observe any clear cases where the binding

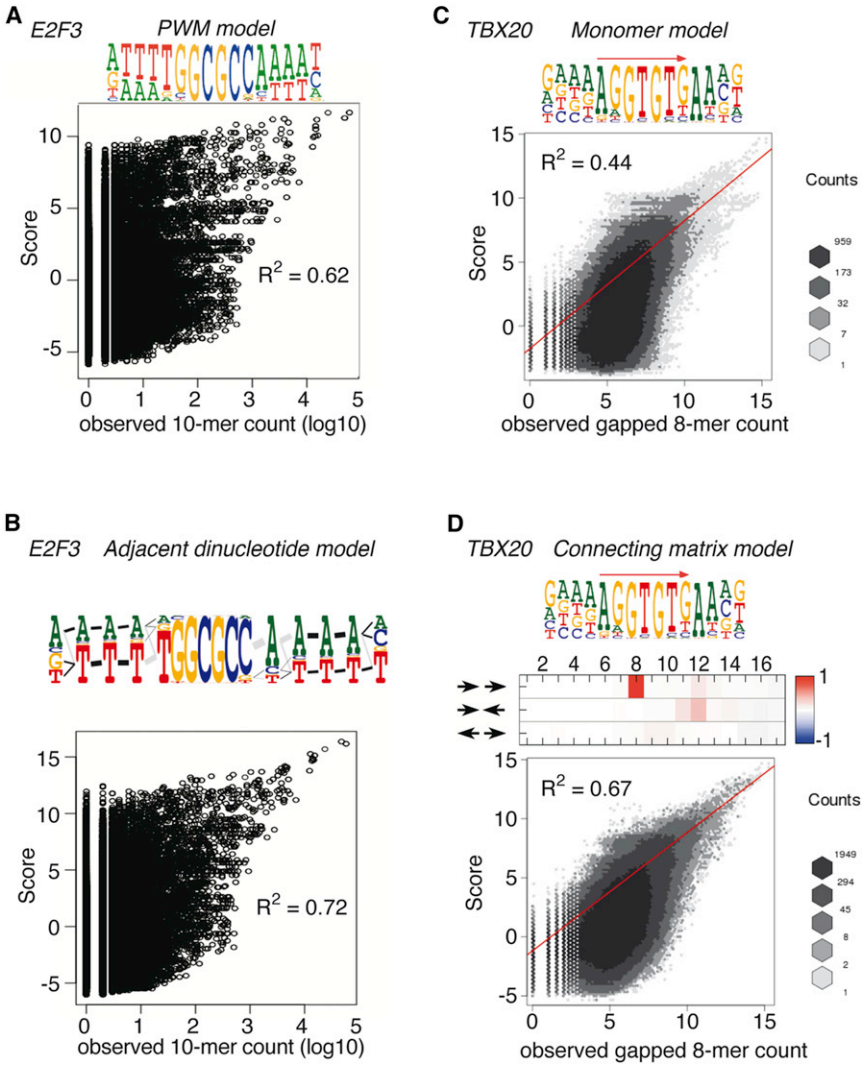

Figure 7. Comparison of Models for TF Binding

(A and B) ADM (B) more accurately describes enrichment of 10 bp subsequences by E2F3 than a conventional PWM (A). In adjacent dinucleotide logo (B), mononucleotide positions that do not explain dinucleotide counts are separated and black edges drawn to indicate the preferred dinucleotides. Gray edges represent dinucleotides that are common but not overrepresented. Thickness of the edges represents the frequency of the indicated dinucleotide; very thin edges are not drawn for clarity.

(C and D) A model consisting of a monomer PWM (canonical monomer target of T box indicated by red arrow) and a spacing and orientation matrix (D) can explain enrichment of gapped 8-mers (4-mer-gap-4-mer) much better than a simple monomer PWM model (C). Heatmap indicates preferred orientations and spacings of the monomers; scale indicates difference in monomer start positions. Red lines indicate least-squares fit; correlation coefficients are also shown. Plots in all panels have logarithmic axes to facilitate visualization; the R^2 values are from the corresponding linear data. See also Table S4.

specificity has changed between human and mouse. However, we did observe several cases where dimer orientation and spacing preferences were divergent between paralogous TFs, suggesting that dimer orientation and spacing preferences evolve faster than primary binding specificities.

Classification of TFs Based on Binding Specificities

Clustering of TFs based on their binding specificities classified them to the known structural families. Many TF families could also be further subclassified based on more subtle differences in specificity within the families (Figure 3) or on a combination of monomer specificity and spacing and orientation preferences. For example, nuclear receptors are known to bind to dimeric sites that vary in both specificity and spacing of the half-sites (Pardee et al., 2011). Clear classification of nuclear receptors to different specificity groups has, however, not been accomplished. The systematic analysis described here allowed classification of nuclear receptors to 12 classes based on a combination of half-site and dimer orientation and spacing preferences. Similarly, although all T box proteins bound to identical half-sites, seven different classes could be identified based on spacing and orientation preferences (Figure 4B). ETS class I proteins also displayed three distinct dimer orientations and spacings (Figure 4A).

A more complex classification of factors was necessary for bZIP proteins, which are known to vary in both specificity and spacing of the half-sites (Amoutzias et al., 2007; Badis et al., 2009; Kim and Struhl, 1995). We find here that many bZIP proteins bind to two sites and that the specificities form a tiled pattern, where in many cases, two factors shared one site and also each bound to another separate site. Such a tiled organization of TF specificity allows a complex control of target genes based on the expression and activity of the particular bZIP factors present in a given cell.

Multiple Binding Modes

The large number of selected sequences, and the large number of factors studied, allowed us also to perform a global genome-wide analysis of common features that are important for recognition of DNA by TFs. It has previously been suggested that many TFs recognize distinctly different sequences (Badis et al., 2009), but this view is controversial (Zhao and Stormo, 2011; Morris et al., 2011). Analysis of our data reveals that multiple PWM models are not needed to explain high-affinity binding of most TFs to DNA. However, multiple binding modes exist for many factors (e.g., bZIP proteins), and most such cases are due to the ability of a factor to bind to both a monomeric and a dimeric site, and/or multiple different dimeric configurations.

Structure-Based DNA Recognition

It is well established that TFs have two primary ways to interact with DNA: a non-sequence-specific interaction with the backbone, and a sequence-specific interaction with the bases (von Hippel and Berg, 1989). The latter is often linked to direct hydrogen bonding between specific DNA bases and DBD amino acids. Most such interactions occur via the major groove of DNA, which is often expanded by an insertion of a DBD recognition helix or loop. A third type of binding that depends on DNA minor groove shape and confers partial sequence specificity has been suggested based on analysis of crystal structures of protein-DNA complexes (Aggarwal et al., 1988; Rohs et al., 2009; Zheng et al., 1999). We find here that such interactions are indeed very common in different TF families and determine their effects on DNA-binding specificity for the first time.

The common DNA structure-based-binding motif is characterized by a core-binding sequence of a TF being flanked by a stretch of either A or T bases. Such interactions are potentially important in formation of consecutive TF-binding sites in regulatory elements. Because this type of recognition of DNA is based on DNA shape, it is also likely that the base preferences of TFs in these regions can be affected by DNA shape changes induced by binding of multiple TFs in close proximity to each other (see also Slattery et al., 2011). Furthermore, due to the fact that TFs can read the minor groove without opening the DNA, such interactions may also increase speed by which TFs locate their target sites (see Elf et al., 2007).

Posterior Homeodomain Proteins and CDX

We also identified another type of correlation between bases that was due to recognition of DNA without hydrogen bonding. All posterior homeodomain proteins (HOX9–HOX13) bound to two types of sites in a partially overlapping pattern. These sites could not be adequately described by a single PWM. Specificities between paralogous HOX proteins (e.g., HOXA13, HOXB13, HOXC13, and HOXD13) were similar to each other, but clear differences were observed between each of HOXs 9, 10, 11, 12, and 13. These differences, combined with proposed latent differences in anterior HOX specificity (Slattery et al., 2011), potentially explain the differences in target specificity of the collinear HOX series.

Interestingly, the parahox CDX proteins that are evolutionarily related to posterior HOX proteins bound to only one type of site that was shared by HOX9 and HOX10 (Figure 6A), suggesting that a partial overlap between the bound sequences has been specifically selected for. Such a partial overlap is also observed between zinc fingers and other TFs, and many TFs in the bZIP family (see above), suggesting that such an arrangement is a common feature of human transcriptional networks.

Role of Base-Stacking Interactions in TF-DNA Binding

In addition to large deviations from the PWM model described above, the large number of sequences analyzed allowed us to identify a general tendency of adjacent bases to affect each other (Figure 5C). The effect of dinucleotide composition on DNA structure is well established (Geggier and Vologodskii, 2010), and dinucleotides are commonly used to predict a large number of properties of DNA, including geometry of the base pairs and melting temperature (Zheng et al., 2010). No clear preference toward or away from any given dinucleotide was found (data not shown), suggesting that TFs do not have a general preference toward a particular structural feature.

Our results indicate that although the primary interactions between TF and DNA occur between individual bases and amino acids, and that independent binding of DNA bases by TFs is generally a good approximation (Benos et al., 2002; Roulet et al., 2002), adjacent bases deviate from this assumption in a manner that is important for quantitative analyses of TF-DNA binding. Thus, in addition to determining base pair geometry and structural features of DNA, adjacent dinucleotides play a role in DNA recognition by TFs. Our results also suggest that systems-biological models of TF-DNA binding based on dinucleotides should perform better in prediction of occupied TF sites than models based on conventional PWMs.

Computational Modeling of Binding

The binding of TFs to DNA is commonly modeled based on a PWM that assumes independence of binding of protein to individual bases. Several alternative models that do not make this independence assumption and, instead, use a larger set of parameters to describe TF-DNA binding have been developed (see, for example, Agius et al., 2010; Roulet et al., 2000; Sharon et al., 2008). Based on our observation that adjacent bases commonly affect each other, and that many TFs bind DNA as monomers or dimers, we developed here two models for TF binding that incorporate these features. The first model is a simple replacement for a PWM that is based on a first-order Markov chain. This model takes into account the effect of adjacent bases and models binding of factors that bind to A or T stretches significantly better than a conventional PWM.

The second model we developed takes into account the spacing and orientation preferences of dimeric sites. This improves models for TFs that bind to DNA both as monomers and dimers or as multiple different dimers. This model can be generalized to heterodimers and chains of TFs of arbitrary type.

The advances in modeling TF-DNA interactions, together with the systematic resource of human TF specificities we describe here, will enable building of more accurate systems-biological models of TF-DNA binding and transcription, thus representing a major step toward decoding of the second, regulatory, genetic code—the code that determines gene expression based on genomic sequences.

EXPERIMENTAL PROCEDURES

Cell Culture, Constructs, and Protein Expression

Human LoVo colon carcinoma and human embryonic kidney-derived 293T (ATCC; CRL-11268) and 293FT cells (Invitrogen; R700-07) were cultured in DMEM with 10% FBS and antibiotics.

Collection consisting of 984 human full-length TFs and 891 DBDs was cloned by PCR from Mammalian gene collection, ORFeome, Megaman cDNA library, or by gene synthesis (Table S1). Another collection composed of 444 mouse DBDs was generated by PCR from templates described earlier by Badis et al. (2009) and Berger et al. (2008). Constructs were sequenced using capillary sequencing (National Public Health Institute, Finland, and MWG, Germany).

For protein production, cells were transfected in 6-well plates using polyethyleneimine (25 kDa; Sigma-Aldrich) with cDNAs in pDEST40-Gau-SBP (Jolma

et al., 2010) or pcDNA3.1-3xFLAG, followed by culture for 2 days and lysis in 1% Triton X-100, 150 mM NaCl, 50 mM Tris-Cl (pH 7.5) with protease inhibitors (cOmplete EDTA-free; Roche). Cell lysates were either deep-frozen at −80°C or used directly. Expression levels of proteins were monitored by luminescence (Renilla Luc assay, Promega; EnVision, PerkinElmer). A subset of 17 and 2 DBDs was expressed as N-terminal thioredoxin-hexahistidine or GST fusions using *E. coli*, respectively (see Extended Experimental Procedures; Table S1).

ChIP-Seq

ChIP-seq for MAFG (antibody: Santa Cruz Biotechnology; sc-22831 X), MAFK (Abcam; ab50322), GMEB2 (Abcam; ab50592), GRHL1 (Abcam; ab77762), HNF1A (Santa Cruz Biotechnology; sc-22840 X), p53 (Santa Cruz Biotechnology; sc-135773 X), HNF4A (Santa Cruz Biotechnology; sc-8987), and E2F7 (Santa Cruz Biotechnology; sc-66870 X) was performed essentially as described in Tuupanen et al. (2009) and J.Y., M.E., and J. Taipale, unpublished data. After sequencing (Illumina GAIIX or Hiseq2000), 4 bp index sequences were removed, and the remaining 33 bp sequences were mapped to the hg18 human reference genome using BWA: mapping quality threshold 20; 3′ bases were trimmed (quality score threshold 20). Duplicate reads were removed to exclude artifactual peaks and to limit PCR bias. Peaks were called using MACS (Zhang et al., 2008), and the motifs generated using MEME, using 61 bp sequences centered on the 500 most enriched peaks (parameters: -revcomp -dna -minw 5 -maxw 20).

HT-SELEX

Detailed SELEX protocol and data analysis are presented in Extended Experimental Procedures. Plate-based HT-SELEX was performed essentially as described in Jolma et al. (2010), except that 14, 20, 30, or 40 bp randomized regions were used. For *E. coli*-produced proteins, a bead-based SELEX protocol was used. Selection ligands contained a 5–6 and 0–3 bp bar code before and after the randomized region, respectively (Table S1).

Raw sequencing data (Illumina GAIIX or Hiseq2000) were binned according to bar codes and analyzed using IniMotif for quality control (see Jolma et al., 2010), identification of the most enriched 6–12 bp subsequences, and generation of primary and secondary PWM models. Final PWM models were generated using the multinomial model (Jolma et al., 2010); cycle and seed sequences are indicated in Table S3.

Nucleotide pairs were counted using the same seed that was used to generate the matrices. Seed was matched exactly outside of the nucleotide pair considered, and the instances of each of the 16 nucleotide pairs were counted. The mononucleotide model describing the nucleotide pairs was generated from the pair counts, and expected nucleotide pair counts were then predicted from this model. The adjacent dinucleotide Markov model (Table S4) was generated by normalizing adjacent nucleotide pair frequencies to generate initial and conditional probabilities.

The connecting matrix model describes the dependence of dimeric binding affinity on spacing and orientation of the two binding sites using a cooperative-binding (cob) table, which has a row for each orientation o (Head-to-Tail, Head-to-Head, and Tail-to-Tail) and a column for each spacing (distance d = 1, 2, …) for a previously obtained monomer PWM. The total score for a dimer site is given as the sum of the PWM scores and the score $cob_{o,d}$ according to the orientation o and spacing d of the two binding sites of the dimer.

Coverage and Similarities between Binding Specificities

To assess the coverage of the model collection, we retrieved the number of human high-confidence TFs (category A; Vaquerizas et al., 2009) that have one or more motifs (HT-SELEX, UniProbe, or JASPAR) for the given TF or a closely related TF (sequence identity = 1 and similarity >0.9, respectively).

The difference between DBD and full-length protein-derived PWMs was analyzed using KL distance (Wei et al., 2010) and compared to replicate experiments for six DBDs (TFAP2A, HES5, ESRRA, CREB3L1, ELK1, HOXD12) from different TF families. For comparison between all profiles, we used SSTAT (Pape et al., 2008), which differentiates better between monomers and dimers. A minimum dominating set, consisting of 239 PWMs, was found by transforming the problem into an integer linear-programming problem, which was then solved optimally using GLPK LP/MIP solver, v.4.43. Detailed computational methods are described in Extended Experimental Procedures.

ACCESSION NUMBERS

Sequencing data has been deposited to ENA under accession numbers ERP001824 and ERP001826.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, four figures, and four tables and one interactive data file can be found with this article online at http://dx.doi.org/10.1016/j.cell.2012.12.009.

ACKNOWLEDGMENTS

We apologize to the authors of original work not cited. Due to the broad scope of this work, and editorial limits on citations, we were largely limited to citing review articles, databases, and more recent original articles not yet reviewed or incorporated into databases. We thank Drs. M. Taipale, M.O. Lombardia, and B. Schmierer for critical review of the manuscript, and R. Nurmi, S. Miettinen, A. Zetterlund, S. Talukder, G. Breard, A. Yang, A. Cote, and H. Zheng for technical assistance. This project was supported by Academy of Finland, Knut and Alice Wallenberg Foundation, Vetenskapsrådet, Cancerfonden, ERC Advanced Grant GROWTHCONTROL, the EU FP7 project SYSCOL to J. Taipale, and CIHR Operating Grant MOP-77721 to T.R.H.

Received: May 25, 2012
Revised: August 18, 2012
Accepted: December 3, 2012
Published: January 17, 2013

REFERENCES

Aggarwal, A.K., Rodgers, D.W., Drottar, M., Ptashne, M., and Harrison, S.C. (1988). Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. Science *242*, 899–907.

Agius, P., Arvey, A., Chang, W., Noble, W.S., and Leslie, C. (2010). High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. PLoS Comput. Biol. *6*, e1000916.

Amoutzias, G.D., Veron, A.S., Weiner, J., 3rd, Robinson-Rechavi, M., Bornberg-Bauer, E., Oliver, S.G., and Robertson, D.L. (2007). One billion years of bZIP transcription factor evolution: conservation and change in dimerization and DNA-binding site specificity. Mol. Biol. Evol. *24*, 827–835.

Babayeva, N.D., Wilder, P.J., Shiina, M., Mino, K., Desler, M., Ogata, K., Rizzino, A., and Tahirov, T.H. (2010). Structural basis of Ets1 cooperative binding to palindromic sequences on stromelysin-1 promoter DNA. Cell Cycle *9*, 3054–3062.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science *324*, 1720–1723.

Balaskas, N., Ribeiro, A., Panovska, J., Dessaud, E., Sasai, N., Page, K.M., Briscoe, J., and Ribes, V. (2012). Gene regulatory logic for reading the Sonic Hedgehog signaling gradient in the vertebrate neural tube. Cell *148*, 273–284.

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.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Peña-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266–1276.

Bohmann, D., Bos, T.J., Admon, A., Nishimura, T., Vogt, P.K., and Tjian, R. (1987). Human proto-oncogene c-jun encodes a DNA binding protein with

structural and functional properties of transcription factor AP-1. Science *238*, 1386–1392.

Brayer, K.J., and Segal, D.J. (2008). Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. Cell Biochem. Biophys. *50*, 111–131.

Brown, R.S. (2005). Zinc finger proteins: getting a grip on RNA. Curr. Opin. Struct. Biol. *15*, 94–98.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106–1117.

Chia, N.Y., Chan, Y.S., Feng, B., Lu, X., Orlov, Y.L., Moreau, D., Kumar, P., Yang, L., Jiang, J., Lau, M.S., et al. (2010). A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature *468*, 316–320.

Davidson, E.H., and Levine, M.S. (2008). Properties of developmental gene regulatory networks. Proc. Natl. Acad. Sci. USA *105*, 20063–20066.

Elf, J., Li, G.W., and Xie, X.S. (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. Science *316*, 1191–1194.

Geggier, S., and Vologodskii, A. (2010). Sequence dependence of DNA bending rigidity. Proc. Natl. Acad. Sci. USA *107*, 15421–15426.

Giguère, V., McBroom, L.D., and Flock, G. (1995). Determinants of target gene specificity for ROR alpha 1: monomeric DNA binding by an orphan nuclear receptor. Mol. Cell. Biol. *15*, 2517–2526.

Grove, C.A., De Masi, F., Barrasa, M.I., Newburger, D.E., Alkema, M.J., Bulyk, M.L., and Walhout, A.J. (2009). A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. Cell *138*, 314–327.

Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., and Taipale, J. (2006). Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell *124*, 47–59.

Jolma, A., and Taipale, J. (2011). Methods for analysis of transcription factor dna-binding specificity in vitro. Subcell. Biochem. *52*, 155–173.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpää, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. *20*, 861–873.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., and Segal, E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362–366.

Kim, J., and Struhl, K. (1995). Determinants of half-site spacing preferences that distinguish AP-1 and ATF/CREB bZIP domains. Nucleic Acids Res. *23*, 2531–2537.

Lamber, E.P., Vanhille, L., Textor, L.C., Kachalova, G.S., Sieweke, M.H., and Wilmanns, M. (2008). Regulation of the transcription factor Ets-1 by DNA-mediated homo-dimerization. EMBO J. *27*, 2006–2017.

Morris, Q., Bulyk, M.L., and Hughes, T.R. (2011). Jury remains out on simple models of transcription factor specificity. Nat. Biotechnol. *29*, 483–484.

Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and Wolfe, S.A. (2008). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell *133*, 1277–1289.

Oliphant, A.R., Brandl, C.J., and Struhl, K. (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. Mol. Cell. Biol. *9*, 2944–2949.

Pape, U.J., Rahmann, S., and Vingron, M. (2008). Natural similarity measures between position frequency matrices with an application to clustering. Bioinformatics *24*, 350–357.

Pardee, K., Necakov, A.S., and Krause, H. (2011). Nuclear receptors: small molecule sensors that coordinate growth, metabolism and reproduction. Subcell. Biochem. *52*, 123–153.

Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W., and Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. *38*(Database issue), D105–D110.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248–1253.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. Annu. Rev. Biochem. *79*, 233–269.

Roulet, E., Bucher, P., Schneider, R., Wingender, E., Dusserre, Y., Werner, T., and Mermod, N. (2000). Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites. J. Mol. Biol. *297*, 833–848.

Roulet, E., Busso, S., Camargo, A.A., Simpson, A.J., Mermod, N., and Bucher, P. (2002). High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. Nat. Biotechnol. *20*, 831–835.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., and Gaul, U. (2008). Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. Nature *451*, 535–540.

Sharon, E., Lubliner, S., and Segal, E. (2008). A feature-based approach to modeling protein-DNA interactions. PLoS Comput. Biol. *4*, e1000154.

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

Solano, R., Nieto, C., Avila, J., Cañas, L., Diaz, I., and Paz-Ares, J. (1995). Dual DNA binding specificity of a petal epidermis-specific MYB transcription factor (MYB.Ph3) from *Petunia* hybrida. EMBO J. *14*, 1773–1784.

Stros, M., Launholt, D., and Grasser, K.D. (2007). The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins. Cell. Mol. Life Sci. *64*, 2590–2606.

Struhl, K. (1987). The DNA-binding domains of the jun oncoprotein and the yeast GCN4 transcriptional activator protein are functionally homologous. Cell *50*, 841–846.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science *249*, 505–510.

Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O., Vanharanta, S., Kivioja, T., Björklund, M., Wei, G., Yan, J., Niittymäki, I., et al. (2009). The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat. Genet. *41*, 885–890.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. *10*, 252–263.

von Hippel, P.H., and Berg, O.G. (1989). Facilitated target location in biological systems. J. Biol. Chem. *264*, 675–678.

Walhout, A.J. (2011). Gene-centered regulatory network mapping. Methods Cell Biol. *106*, 271–288.

Wei, G.H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., et al. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J. *29*, 2147–2160.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhao, Y., and Stormo, G.D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat. Biotechnol. *29*, 480–483.

Zheng, G., Colasanti, A.V., Lu, X.J., and Olson, W.K. (2010). 3DNALandscapes: a database for exploring the conformational features of DNA. Nucleic Acids Res. *38*(Database issue), D267–D274.

Zheng, N., Fraenkel, E., Pabo, C.O., and Pavletich, N.P. (1999). Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. Genes Dev. *13*, 666–674.

EXTENDED EXPERIMENTAL PROCEDURES

Clone Design and Protein Expression

Generally, DBD clones included all DBDs present in the corresponding genes (see Table S1 for details). All DBD clones were sequenced from both ends and all full-length clones were confirmed by sequencing from at least one end.

For protein expression, clones in Gateway recombination cloning entry vector were transferred to the mammalian protein production vectors pDEST40-Gau-SBP (Jolma et al., 2010) or pcDNA3.1-3xFLAG, or the bacterial expression vectors pDEST15-MAGIC (Berger et al., 2006), pETG20A (Vincentelli et al., 2011) or pETG20A-SBP, where original pETG20A was modified by synthesis and addition of C-terminally inserting SBP tag (Geneart, Germany for sequence see Table S1).

Mammalian expression is explained in main Experimental Procedures. For bacterial expression, proteins were expressed in Rosetta 2(DE3)pLysS or C41 strains using auto-inducing ZYP5052 medium or IPTG induction, respectively. IPTG induction was performed in 250 ml cultures at 1 mM at 16°C overnight. For autoinduction (Vincentelli et al., 2011), overnight culture of bacteria was inoculated in 1:40 ratio into ZYP5052 medium, and grown for 4 hr at +37°C to followed by protein expression for 18-25 hr at +17°C. Bacteria were collected by centrifugation and lysed, followed by GST or his-tag purification of the proteins with high performance glutathione Sepharose or Ni Sepharose 6 Fast Flow (GE healthcare). Glutathione Sepharose beads were used directly for SELEX. Ni Sepharose purified proteins were eluted with Imidazole and stored in 300 mM NaCl, 125 mM Imidazole, 50% glycerol in 50 mM Tris-Cl, pH 8 at −20°C. Protein expression was verified by SDS-PAGE and Coomassie brilliant blue staining.

SELEX

In HT-SELEX, an excess of double-stranded DNA fragments containing a randomized region (Table S1) were allowed to bind to immobilized TFs. Different selection ligands contained constant sequences on both sides that allowed analysis of proteins that recognize partial sequences on any given ligand. Unbound DNA was removed by rapid washing, after which bound DNA was eluted, amplified and sequenced. This process was repeated multiple times, and aliquots of the enriched DNA were sequenced at each cycle. Sequencing is described in Experimental Procedures. Read length for SELEX experiments was longer than the 5′ barcode, variable region and 3′ barcode combined.

Plate-based SELEX was performed as described in Jolma et al. (2010), except for the use of a liquid handling workstation (Agilent Bravo) to perform the assays, using a 7 min wash protocol composed of 25 wash steps with two consecutive BioTek405 CW plate washers (to avoid contamination). Streptavidin plates (Thermo Scientific 15502) were used for capture of the proteins, except for CTCF, for which anti-Flag epitope antibody coated plates (Sigma-Aldrich, cat nr. P2983) were used.

Bead-based SELEX was performed using glutathione sepharose, Ni Sepharose 6 Fast Flow or streptavidin agarose beads (Thermo, cat 20359; see Table S1 for details). Briefly, 100-250 ng of purified soluble protein or beads containing immobilized protein, 100 ng of selection ligand and 83 ng of non specific poly-dIdC oligonucleotide competitor was incubated for 10 min in 20 μl volume of binding buffer (80 mM NaCl, 37.5 mM Imidazole, 0.7 mM $MgCl_2$, 0.35 mM EDTA, 0.7 mM DTT, 17.8% glycerol in 7 mM Tris-Cl pH 7.5). Subsequently, soluble TFs and bound DNA were captured by incubation with 150 μl of Ni Sepharose 6 Fast Flow beads (equilibrated in 50 mM NaCl, 1 mM $MgCl_2$, 0.5 mM EDTA, 4% Glycerol in 10 mM Tris-Cl, pH 7.5) for 20 min with constant shaking. Beads were washed by vacuum filtration (Millipore 96-well filter plate MSDVN6550) 12 times with 200 μl of bead equilibration buffer for 5 min each. Beads were suspended to 200 μl water and stored at −20°C or used directly as a template for PCR.

SELEX Data Analysis

To model TF binding, we generated position weight matrices using a multinomial method that yields profiles that are similar to those generated using maximum likelihood methods such as BEEML (Zhao et al., 2009). The multinomial method (Figure 1B) was used because it considers only closely related sequences, and thus allows simple identification of cases where multiple models are required to explain binding of one TF to DNA. The processing pipeline corrected for non-specific DNA carryover, and resulted in profiles that were similar to those obtained using ratio-based models.

IniMotif was used to identify experiments that displayed enrichment of specific sequences. Primary model used the most enriched k-mer as seed, and the secondary model used the most enriched k-mer from the sequence reads that were not used to generate the primary model. These initial matrices were inspected to rule out problems such as complexity bottlenecks, DNA contamination or biases due to binding of factors to specific barcode sequences (see Jolma et al., 2010), and to identify experiments where specific sequences were enriched. After the initial IniMotif analysis, all sequences for all cycles for samples that displayed robust enrichment of specific sequences were loaded to a MySQL database together with information about the experiments and the factors analyzed.

In general, cycle used for PWM generation was selected from cycles 2 to 4 in such a way that more than 1000 subsequences were included to the model after background correction, and the seed length and sequence was selected based on the following criteria: The length of seed sequence was determined by including flanking positions where the ratio between the most and least frequent bases was > 2. After initial PWM generation using the most frequent kmer as seed, the seed was made more redundant to accommodate more sequences at positions where the frequency of the most common base was < 0.5. At these positions, we used either N, or where the ratio between the second and third most frequent bases was > 2, we used the IUPAC symbol for the two most frequent bases (R, Y, M, K, S, or W).

If the length of the seed was longer than 10 bp, a multinomial model allowing a single mismatch at any position was used. Seed sequences were further manually curated to prevent mixing of two distinct binding modes, and to distinguish between monomer and dimer models. Multiple seeds were used for the same factor if the IniMotif analysis of 6 to 12-mer sequences, or if plotting of the observed k-mers versus those expected from the PWM revealed that the first model did not explain the most enriched k-mers, or if enrichment of dimers was observed.

Models were corrected for background by subtracting normalized counts from the previous round as described in (Jolma et al., 2010) using the equation $M_{corrected} = M_{k+1}$ - lambda * M_k, where lambda is the fraction of DNA carried over non-specifically estimated using 8-mer frequencies, and M_{k+1} and M_k are the uncorrected matrices normalized for number of input sequences from cycles k+1 and k, respectively. Enrichment ratio model (Figure 1C) was calculated according to the following equation: M_{ratio} = (M_{k+1}+ pseudocount)/(M_k+pseudocount)-lambda. A pseudocount at 1% of maximum frequency at cycle k+1 was used to prevent spurious results due to division with small numbers.

We also investigated whether the MEME expectation maximization algorithm that builds models using a larger sequence space than what we used would result in models that better describe the enrichment of k-mers. In general, this was not found to be the case (not shown; see also Jolma et al., 2010). Use of the multinomial method was selected because: 1) it exactly corresponds to the PWM model representation, 2) the number of enriched sequences was very high, alleviating the need to analyze a large sequence space, and 3) use of larger sequence space commonly resulted in mixing of multiple different models.

The resulting data set contains information for more TFs than the entire published literature. The data were generated using a systematic expression and SELEX pipeline, and the process was run using an automated platform generating highly consistent and intercomparable data. The complete resource including all quantitative data is included as a Supplement, and primary sequence data is available on Short Read Archive.

Adjacent Dinucleotide and Connecting Matrix Models

For generation of adjacent dinucleotide models, all subsequences within Hamming distance of two from a consensus seed sequence (seed) were identified, and nucleotide pair counts were generated by counting the instances of each nucleotide pair at given pair of positions when all other bases exactly matched the seed. Initial probability for a given base was then calculated by adding the probabilities for all dinucleotides starting with that base. This was performed at all positions to allow scoring of kmers that were shorter than the matrix. The conditional probabilities were then calculated as follows: For each dinucleotide starting with A, C, G, or T, the sum of the probabilities for all possible second bases was normalized to one. For scoring, all matrices were converted to log odds form.

Connecting matrix model was generated as described in Experimental Procedures. Positive and negative values in the cob table denote the strength of preference and rejection of the corresponding dimers, respectively. By using the number of observed dimers and the number of expected dimers in a background model, we define each value in the cob table as follows: $cob_{o,d} = \log_2(observed_{o,d}/expected_{o,d})$. The background model consists of probabilities P_x and P_y for all nucleotide sequences x and y of length m and d, respectively. These background probabilities are estimated from the data as the relative frequencies.

Here the quantity $observed_{o,d}$ is evaluated by counting each binding site pair in the data, where both sites have PWM score above a fixed threshold t. The expected dimers are obtained from the background probabilities as follows: $expected_{o,d} = n^*P_{dimer(o,d)}$, where n is the number of windows of length m+d in the data and $P_{dimer(o,d)}$ is the probability of finding the dimer in a random sequence of length m+d. In order to estimate the dimer probability, we define three sets of sequences. If PWMs M1 and M2 are directed according to the orientation o, we define sets T1 and T2 of sequences that get score higher than the threshold t with PWMs M1 and M2, respectively. The set S of dimeric sequences consists of those sequences a of length m+d that have binding site of M1 as the prefix, and binding site of M2 as the suffix of a.

If d>=m, that is, the sites are not overlapping, we get the probability $P_{dimer(o,d)}$ from the product $P_{T1}{}^*P_{T2}$ of the background probabilities. If, on the other hand, d < m, we get the probability $P_{dimer(o,d)}$ as the sum of terms $\frac{1}{2}P_b{}^*p_c + \frac{1}{2}P_{c'}{}^*P_{b'}$ over all sequences a = bc = c'b' in S, where the subsequences b and b' have length m, and subsequences c and c' have length d.

The total score for a dimer site is given as the sum of the PWM scores and the score $cob_{o,d}$ according to the orientation o and spacing d of the two binding sites of the dimer.

ChIP-Seq and ROC Analysis

Antibodies and references for ChIP-seq are indicated in Experimental Procedures. Briefly, LoVo cells were crosslinked by 1% formaldehyde and chromatin was sheared to 200-500 bp fragments, and immunoprecipitated with 5 μg of specific antibodies or non-specific IgG. After washing, TF complexes were extracted, treated with RNase A (Dnase free, Sigma-Aldrich Cat.No. R6513) followed by proteinase K (Thermo Scientific Cat.No. EO0491) at 65°C degrees to reverse the cross-links and to digest proteins. DNA was purified using QIAGEN PCR purification kit and libraries for massively parallel sequencing constructed as described (J.Y. et al., unpublished data).

For receiver operating characteristic (ROC) curves of a PWM in a ChIP-seq experiment, we created a positive set of DNA sequences by extracting the 250bp of genomic DNA (hg18 assembly) flanking the peak summits of the 500 ChIP-seq peaks with the lowest p-values. As a negative set we used same sized sequences flanking random mappable genomic positions. The best scoring match to the relevant PWM was recorded for each sequence, and for each score threshold the fraction of sequences in the positive set with a match above that threshold (true positives) was plotted against the corresponding fraction in the negative set (false positives).

Analysis of Coverage of the TF Collection

To determine the coverage of the TF collection and to compare it to existing data, we ran pfam_scan.pl on all protein sequences for the genes corresponding to *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* motifs from our data, UniProbe PBM database, and JASPAR. We then extracted the protein sequences corresponding to the predicted DBD regions and ran pairwise sequence alignment of all pairs of such protein sequences that correspond to the same pfam class. For alignment, we used the Needleman-Wunsch alignment algorithm with the BLOSUM62 scoring matrix and default parameters. We created a relational database containing the resulting sequence similarities linked to the motifs via protein sequences and corresponding gene IDs.

Comparison of Similarity between the PWM Models

We compared the PWMs to each other using three different methods: the minimum Kullback-Leibler divergence-based method, a method based on ungapped alignment of motifs (TOMTOM; Bailey et al., 2009), and a method based on covariance of binding site positions in random DNA sequence (SSTAT; Pape et al., 2008). All analyses yielded similar overall pattern of similarity between the matrices. However, differences were observed in distances obtained for dimeric sites.

Neither of the two widely used measures that compare the probability distributions defined by two PWMs, Pearson correlation coefficient (TOMTOM software; Gupta et al., 2007) and the Kullback-Leibler distance could clearly separate different dimeric binding modes of TFs with similar monomer specificities. In such a case, even though there is a clear resemblance between the dimer PWMs (a shared half-site), the dimers may totally lack overlapping high-affinity binding sites (for example when the monomers are in opposite orientations within the dimers). However, SSTAT could separate these cases because it explicitly regards two PWMs to be similar if they describe similar sets of binding sites. The similarity of the binding sites is defined as the covariance of the number of binding sites the two PWMs have on a random DNA sequence. Thus, the covariance-based method was selected for analyses of the complete collection, as it more clearly separated dimers and monomers containing similar subsequences.

Network Analysis of Similarity between the PWMs

To visualize the data, we first calculated the similarities of all pairs of 830 PWMs using SSTAT (parameters: 50% GC-content, pseudocount regularization, type I threshold 0.01). These settings give the same overall density of binding site occurrences for both PWMs tested, and limit the effects of stringency and low affinity sites on the similarity score.

We then generated a network containing two types of nodes, one type representing TF binding profiles, and another type representing TF proteins.

TF protein nodes were connected to their binding models, and the binding models were further connected to each other if their SSTAT similarity score (asymptotic covariance) was greater than 1.5×10^{-5}. This cut-off was determined by visual inspection of the connected and unconnected PWMs and resulted in a network with 3563 edges between PWMs.

Minimum dominating set of the network (Garey, Michael R.; Johnson, David S. (1979), Computers and Intractability: A Guide to the Theory of NP-Completeness, W. H. Freeman) was used to select the representative set of the PWMs. A dominating set of the PWM network covers all specificities, as defined by the similarity measure and its cut-off value, because every other PWM node is connected by an edge to at least one PWM node in the dominating set. The minimum dominating set is the smallest possible such set and is thus a concise representation of the binding specificities.

Finally, the network was visualized using Cytoscape software v.2.8.0 (Smoot et al., 2011). The layout of the network was done using yFiles organic algorithm. Networks were transferred to vector image editor, and further annotated and labeled; 11 matrices having very low sequence counts, or that represented replicate experiments and/or technical controls that are indicated in orange in Table S3 were removed from final images.

For comparison, we also constructed networks containing PWMs downloaded from JASPAR database (downloaded files last updated October 12, 2009; Portales-Casamar et al., 2010). We took all PWMs which species annotation included human or mouse or both except those in the collection CNE as they cannot be directly linked to individual TFs. When there were several versions available for the same base identifier, only the newest one was chosen. The resulting set of 500 JASPAR PWMs were included into the networks that were otherwise constructed the same way as the one containing only SELEX PWMs.

K-mer Network Analysis

For visualization of subsequences enriched in the fourth cycle of CDX1, CDX2, Hoxc10, HOXC11, HOXC12, HOXB13 and Hoxd9 experiments (Figure 6A) in cytoscape, we generated a network describing the relationships between the factors and the 9-mers, we selected all 9-mers that were in the top five enriched 9-mers from any of the experiments, and ended in the sequence "TAAAA." Edges were then drawn between the factors and the 9-mers if a 9-mer was enriched in a given experiment. To represent the similarity between the sequences, an edge was also drawn between the 9-mers if they were within a Hamming distance of one from each other.

SUPPLEMENTAL REFERENCES

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*(Web Server issue), W202–W208.

Bich, C., Bovet, C., Rochel, N., Peluso-Iltis, C., Panagiotidis, A., Nazabal, A., Moras, D., and Zenobi, R. (2010). Detection of nucleic acid-nuclear hormone receptor complexes with mass spectrometry. J. Am. Soc. Mass Spectrom. *21*, 635–645.

Conlon, F.L., Fairclough, L., Price, B.M., Casey, E.S., and Smith, J.C. (2001). Determinants of T box protein specificity. Development *128*, 3749–3758.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol. *8*, R24.

Lin, R., Génin, P., Mamane, Y., and Hiscott, J. (2000). Selective DNA binding and association with the CREB binding protein coactivator contribute to differential activation of alpha/beta interferon genes by interferon regulatory factors 3 and 7. Mol. Cell. Biol. *20*, 6342–6353.

Mader, S., Leroy, P., Chen, J.Y., and Chambon, P. (1993). Multiple parameters control the selectivity of nuclear receptors for their response elements. Selectivity and promiscuity in response element recognition by retinoic acid receptors and retinoid X receptors. J. Biol. Chem. *268*, 591–600.

Mohibullah, N., Donner, A., Ippolito, J.A., and Williams, T. (1999). SELEX and missing phosphate contact analyses reveal flexibility within the AP-2[alpha] protein: DNA binding complex. Nucleic Acids Res. *27*, 2760–2769.

Roche, P.J., Hoare, S.A., and Parker, M.G. (1992). A consensus DNA-binding site for the androgen receptor. Mol. Endocrinol. *6*, 2229–2235.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics *27*, 431–432.

Stroud, J.C., Lopez-Rodriguez, C., Rao, A., and Chen, L. (2002). Structure of a TonEBP-DNA complex reveals DNA encircled by a transcription factor. Nat. Struct. Biol. *9*, 90–94.

Vincentelli, R., Cimino, A., Geerlof, A., Kubo, A., Satou, Y., and Cambillau, C. (2011). High-throughput protein expression screening and purification in *Escherichia coli*. Methods *55*, 65–72.

Welboren, W.J., van Driel, M.A., Janssen-Megens, E.M., van Heeringen, S.J., Sweep, F.C., Span, P.N., and Stunnenberg, H.G. (2009). ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands. EMBO J. *28*, 1418–1428.

Zhao, Y., Granas, D., and Stormo, G.D. (2009). Inferring binding energies from selected binding sites. PLoS Comput. Biol. *5*, e1000590.

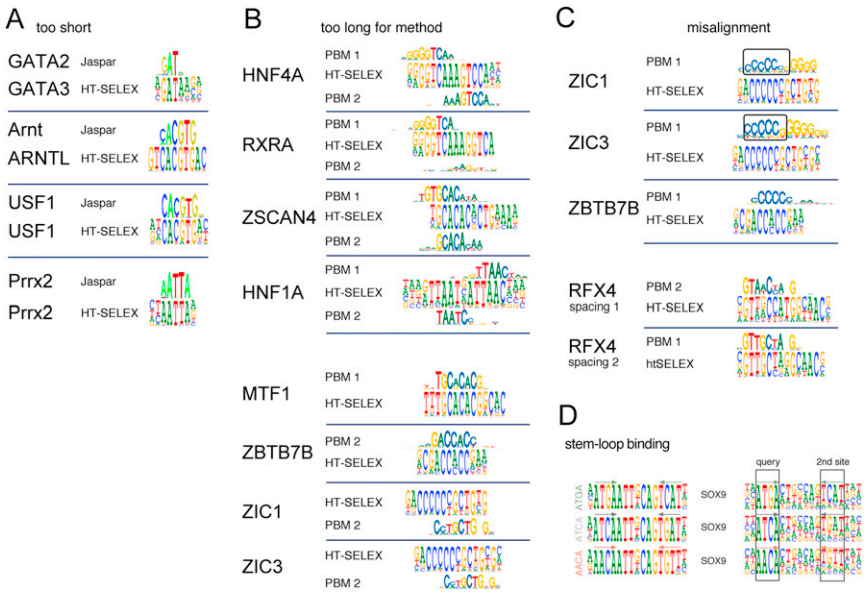

Figure S1. Differences between HT-SELEX Data and Existing Data, Related to Figure 2

(A) Some differences can be explained by the low information content of previous models compared to HT-SELEX data for the same factor or a close paralog. Name of factor, source of model and sequence logo are shown.

(B) In some cases, HT-SELEX identified motif is longer than what the previous method could analyze. The previous models have broken into two separate models (PBM1 and PBM2; top), or is only partially represented by previous data (bottom).

(C) Misalignment in PWM generation from kmer data can result in formation of false palindromic sites (ZIC, ZBTB7B, top; misaligned sequence boxed), or inappropriate joining of two parts of two distinct models for the same factor (RFX4, bottom).

(D) HT-SELEX detects SOX binding to inverted repeat sequences that apparently represent a stem-looped single-stranded DNA. Left: three different apparent dimers are bound by SOX9. Right: sequences flanking ATGA (top), ATCA (middle) and AACA (bottom) query matches reveal that in each case, an inverted repeat of the query sequence appears 3′ to the query after a 7 bp gap (2nd site), suggesting that the bound sequence is a stem-loop formed from a single-stranded DNA. This interpretation is also consistent with the preferential presence of such matches in only one strand of the selection ligand (not shown).

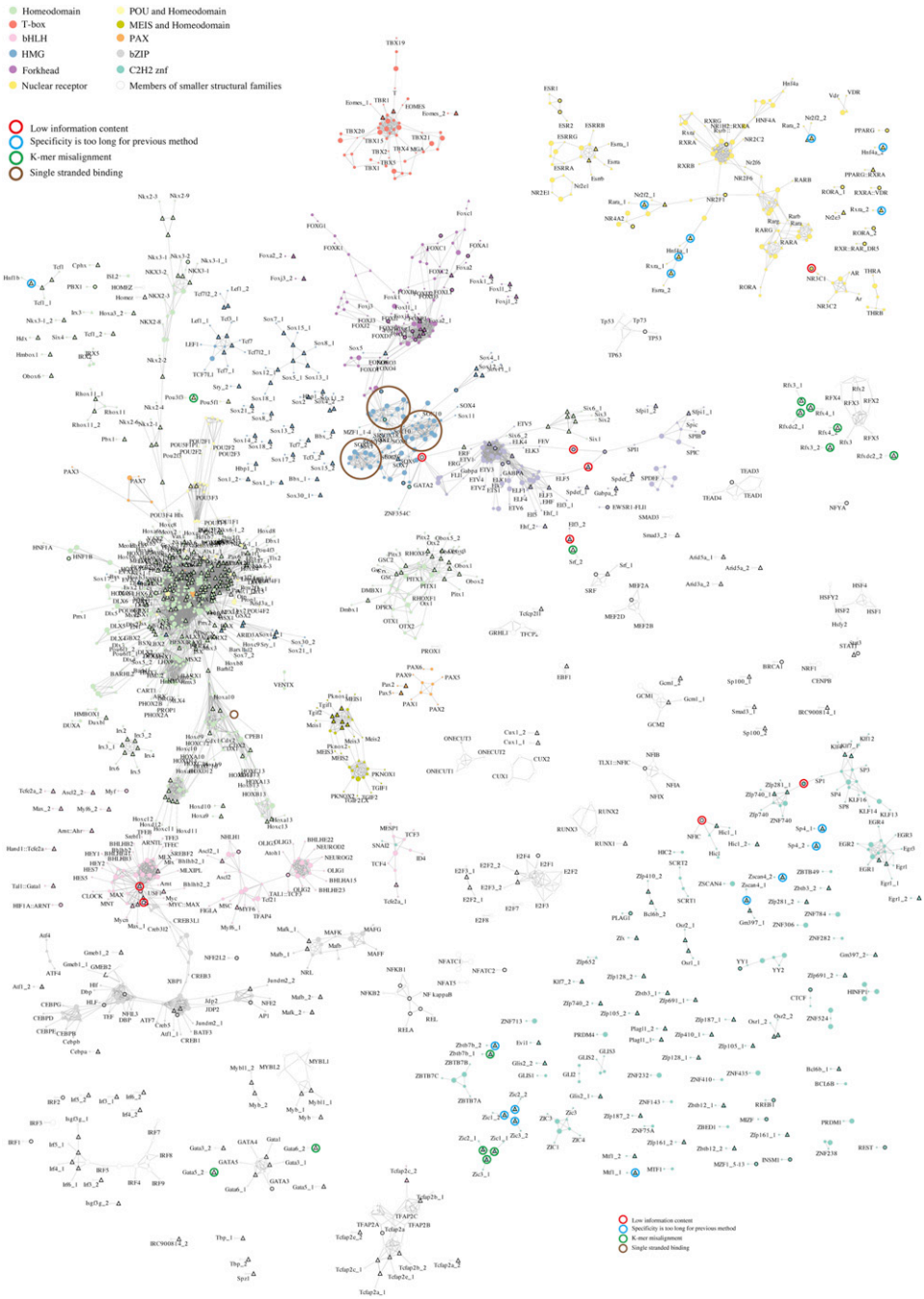

Figure S2. Network Representation of HT-SELEX, JASPAR CORE, and PBM Data, Related to Figure 3

SELEX models have thin edge, whereas models derived from JASPAR or PBM are indicated with a thick edge. Network is laid out as in Figure 3. Individual binding models are colored according to the TF family indicated (top left). Some profiles that diverge between SELEX and JASPAR data are indicated by colored circles; red indicates that motif is of low information content, blue that HT-SELEX motif is longer than the previous method could analyze, green that motif is affected by misalignment of kmers or other computational processing artifacts, and brown that motif represents sequences potentially enriched as single-stranded or stem-looped DNA in HT-SELEX. Detailed analysis of each type case is shown in Figure S2. Analysis of mouse TF specificities using PBMs has revealed that TFs can have multiple modes of DNA binding, and that a single TF can bind to distinctly different sequences (Badis et al., 2009), suggesting that PWM models might not adequately describe DNA-binding for most TFs. The two types of PBM models are indicated by an underscore followed by number (1 is primary, 2 secondary). Note that most PBM models that are connected to SELEX models are primary, whereas most secondary models from PBM do not connect to SELEX motifs.

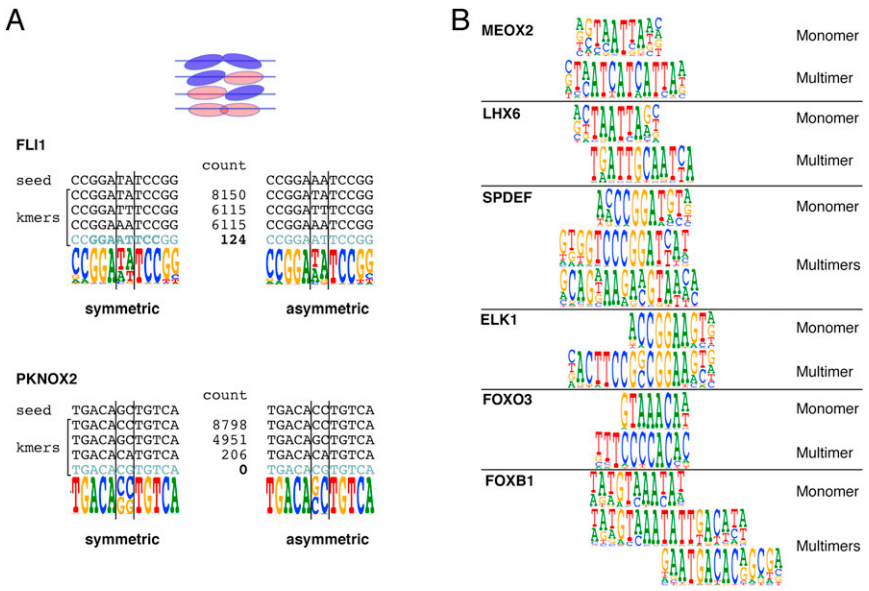

Figure S3. Closely Packed Monomers Can Affect Each Other's Binding Specificity, Related to Figure 6

(A) Asymmetric binding of the monomers is observed when monomer sites are located close together. Top: Close packing of target sites can affect monomer specificity. Protein can bind to an optimal site (pink oval) or to a weaker site (blue oval). Middle: The consensus sequence of FLI1 dimer expected by monomer specificity GGAATTCC (bottom, gray) is very weakly enriched, whereas sites where one or both monomers bind to a GGAT core are strongly enriched. Note that the asymmetric PWM (right) correctly describes lack of enrichment of the GGAATTCC site, whereas the symmetric PWM (left) predicts much higher enrichment for this sequence. Bottom: Similar effect is observed in a PKNOX2 dimer.

(B) Potential cases where dimerization or multimerization affects monomer specificity even more dramatically.

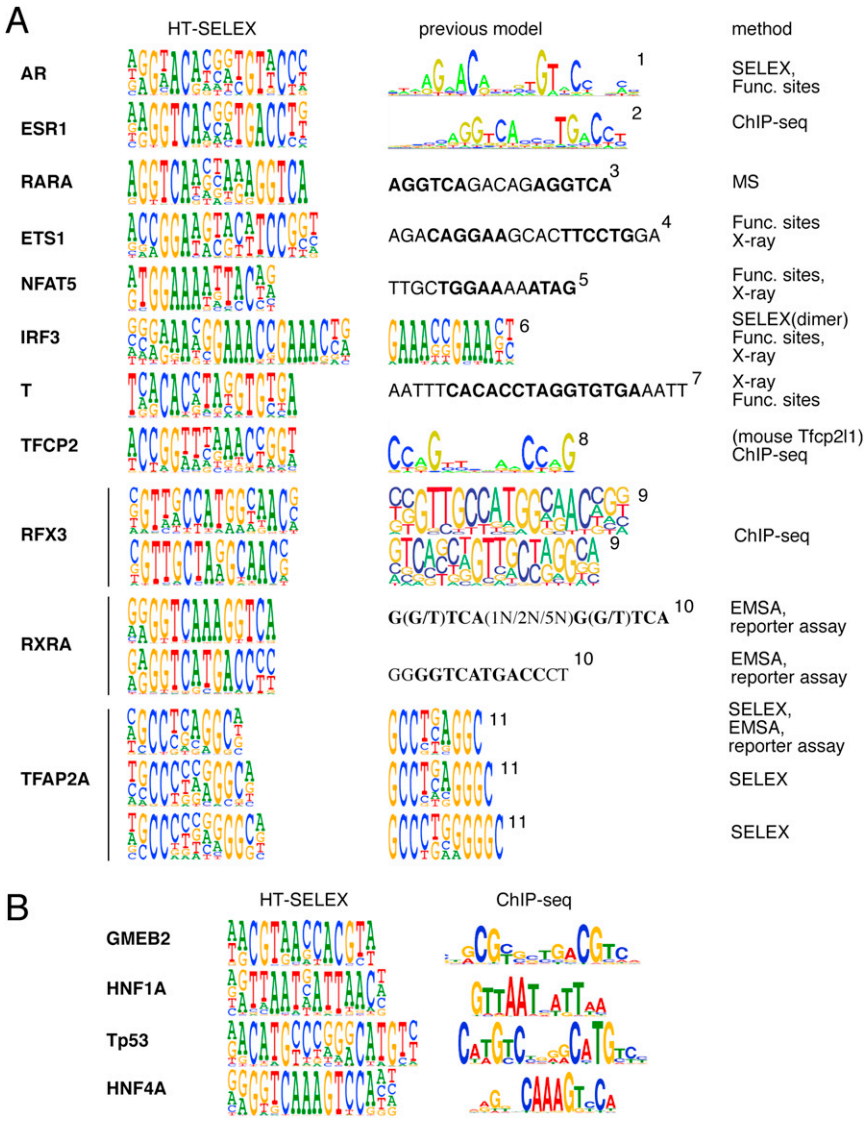

Figure S4. Literature and ChIP-Seq-Based Validation of Homodimeric Interactions Detected in HT-SELEX, Related to Figure 1

(A) Comparison between binding models obtained using SELEX (left panel) and previously published models (right panel). References for previous models: [1]Roche et al., 1992; [2]Welboren et al., 2009; [3]Bich et al., 2010; [4]Lamber et al., 2008; [5]Stroud et al., 2002; [6]Lin et al., 2000; [7]Conlon et al., 2001; [8]Chen et al., 2008; [9]Jolma et al., 2010; [10]Mader et al., 1993; [11]Mohibullah et al., 1999. Previous models are based on X-ray structure (X-ray), SELEX, ChIP-seq, mass spectrometry (MS), reporter assay, electrophoretic mobility shift assay (EMSA) or known functional sites (Func. sites).

(B) *In-vivo* confirmation of homodimeric binding models for GMEB2, HNF1A, Tp53 and HNF4A. ChIP-seq peaks from where the enriched motifs were identified are from Yan et al., in preparation.