Diversity and Complexity in DNA Recognition by Transcription Factors

Gwenael Badis,[1]* Michael F. Berger,[2,3]* Anthony A. Philippakis,[2,3,4]* Shaheynoor Talukder,[1,5]* Andrew R. Gehrke,[2]* Savina A. Jaeger,[2]* Esther T. Chan,[5]* Genita Metzler,[6] Anastasia Vedenko,[7] Xiaoyu Chen,[1] Hanna Kuznetsov,[6] Chi-Fong Wang,[8] David Coburn,[1] Daniel E. Newburger,[2] Quaid Morris,[1,5,9,10] Timothy R. Hughes,[1,5,10]† Martha L. Bulyk[2,3,4,11]†

Sequence preferences of DNA binding proteins are a primary mechanism by which cells interpret the genome. Despite the central importance of these proteins in physiology, development, and evolution, comprehensive DNA binding specificities have been determined experimentally for only a few proteins. Here, we used microarrays containing all 10–base pair sequences to examine the binding specificities of 104 distinct mouse DNA binding proteins representing 22 structural classes. Our results reveal a complex landscape of binding, with virtually every protein analyzed possessing unique preferences. Roughly half of the proteins each recognized multiple distinctly different sequence motifs, challenging our molecular understanding of how proteins interact with their DNA binding sites. This complexity in DNA recognition may be important in gene regulation and in the evolution of transcriptional regulatory networks.

The interactions between transcription factors (TFs) and their DNA binding sites are an integral part of the gene regulatory networks that control development, core cellular processes, and responses to environmental perturbations. However, only a handful of sequence-specific TFs have been characterized well enough to identify all the sequences that they can and, just as importantly, cannot bind. Computational analysis of microarray readout of chromatin immunoprecipitation experiments (ChIP-chip) suggests extensive use of low-affinity binding sites in yeast (*1*), and computational models of gene expression during fly embryonic development suggest that low-affinity binding sites contribute as much as high-affinity sites (*2*).

The availability of TF binding data spanning the full affinity range would improve our understanding of the biophysical phenomena underlying protein-DNA recognition and would also improve accuracy in analyzing cis regulatory elements. Here we report the comprehensive determination of the DNA binding specificities of 104 known and predicted mouse TFs with the use of the universal protein binding microarray (PBM) technology (*3*). These TFs represent 22 different DNA binding domain (DBD) structural classes that are the major DBD classes found in metazoan TFs.

We created N-terminal glutathione *S*-transferase (GST) fusion constructs of the DBDs of 104 known and predicted mouse TFs (fig. S1 and table S1) (*4*). Five of these proteins—Max, Bhlhb2, Gata3, Rfx3, and Sox7—were also represented as full-length fusions to N-terminal GST, yielding a total set of 109 nonredundant proteins represented by 115 samples (*5*). Each protein was used in two PBM experiments (*6*, *7*) (figs. S2 to S4 and table S2). DNA binding site motifs were initially derived by the Seed-and-Wobble algorithm (*3*, *8*); Seed-and-Wobble first identifies the single 8-mer (ungapped or gapped) with the greatest PBM enrichment score (E score) (*3*) and then systematically tests the relative preference of each nucleotide variant at each position, both within and outside the seed (*5*). Later analyses incorporated additional motif-finding algorithms, including RankMotif++ (*9*) and Kafal (*5*).

Beyond simply providing a DNA binding site motif, these data provide a rank-ordered listing of the preference of a protein for every gapped and ungapped *k*-mer "word," where *k* is the number of informative nucleotide positions in the binding site. This data set consists of 9.6 million measurements, from which we can derive binding data for 22.3 million ungapped and gapped 8-mers (up to 12 positions) for each protein. For each of the 8-mers for each protein, we report its E score, median signal intensity Z score, and false discovery rate *Q* value (*5*). We found that the average number of ungapped 8-mers considered "bound" at a *Q* value threshold of 0.001 varied across classes, ranging from 65 for the MADS class factor SRF to 871 for the E2F class.

For TFs that had previously known binding site motifs, we observed general agreement with earlier motif data (fig. S5 and table S3) (*5*). Comparisons to dissociation constant data (*10*) for Max and for the yeast TF Cbf1 (*3*) indicate that words with higher E scores are generally bound with higher affinity (*3*) (fig. S6). Confirmation by electrophoretic mobility shift assays (EMSAs) for three newly characterized proteins and one recently characterized protein (*11*)—Zfp740, Osr2, Sp100, and Zfp161 (ZF5) (*12*), respectively—is shown in fig. S7.

To examine correlations among the proteins' DNA binding specificities and to identify DNA sequences that distinguish the binding profiles of different TF families, we hierarchically clustered the *k*-mers that met a stringent binding threshold (E score $\geq$ 0.45) for at least one of the proteins. We used E scores because they are robust to differences in protein concentration and thus facilitate comparison of *k*-mer data across arrays (*3*); we consider them as a proxy for relative affinities. Different DBD classes generally recognize distinct portions of sequence space (Fig. 1A and fig. S8). However, even proteins with up to 67% amino acid sequence identity exhibited distinct DNA binding profiles. For example, although Irf4 and Irf5 both bind the same highest-affinity sites (8-mers containing CGAAAC), they prefer different lower-affinity sites (TGAAAG versus CGAGAC) (Fig. 1B). We verified for five TFs that the full-length protein displays a virtually identical spectrum of 8-mer preferences to that of the DBD and that the spectrum is distinct from other proteins of the same structural class (figs. S2 and S9).

Our data set includes most members of three TF structural classes in mouse: (i) Sox and Sox-related, (ii) IRF, and (iii) AP-2. In an extreme case, we find no evidence that the binding profiles of the AP-2 class members are different from each other (fig. S9B), consistent with reports that the human counterparts of AP-2α, AP-2β, and AP-2γ all bind GCCNNNGGC (*13*). In contrast, members of the IRF class all appeared to have different binding profiles (fig. S9L).

The Sox and Sox-related family presents an intriguing instance of highly conserved DBDs with closely related but distinct binding preferences. We found marked differences in the binding specificities of the Sox (*14*), Tcf/Lef (*15*, *16*), and Hbp1/Bbx (*17*) families (Fig. 1C). In most cases, our data are roughly consistent with known binding sequences (Fig. 1C), although there are also clear differences: Hpb1 and Bbx have been described as preferring WRAATGGG (*17*), whereas in our data, Hbp1 and Bbx prefer TGAATG and have lesser preference for AATGGG. Our data confirm that there are at least four different varieties of Sox and Sox-related DNA binding specificity (Fig. 1C) and suggest that there are subtle variations among Sox proteins (Fig. 1B).

Several TFs had two distinct sets of high-scoring *k*-mers. For example, the nuclear receptor

[1]Banting and Best Department of Medical Research, University of Toronto, Toronto, ON M5S 3E1, Canada. [2]Division of Genetics, Department of Medicine, 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]Harvard–Massachusetts Institute of Technology (MIT) Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA. [5]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada. [6]Department of Biology, MIT, Cambridge, MA 02139, USA. [7]Department of Biology, Wellesley College, Wellesley, MA 02481, USA. [8]Department of Physics, MIT, Cambridge, MA 02139, USA. [9]Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada. [10]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON M5S 3E1, Canada. [11]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: t.hughes@utoronto.ca (T.R.H.); mlbulyk@receptor.med.harvard.edu (M.L.B.)

hepatic nuclear factor 4 alpha [Hnf4a; C4 zinc finger (ZnF) DBD] exhibits strong binding to both sequences containing GGTCCA and sequences containing GGTCCA (Fig. 2A), whereas all four other C4 ZnF TFs that we examined bind only to GGTCA. We confirmed binding of Hnf4a to both variants by EMSA (fig. S10). TFs that can recognize two distinctly different DNA sequences have been noted before (*18*). We

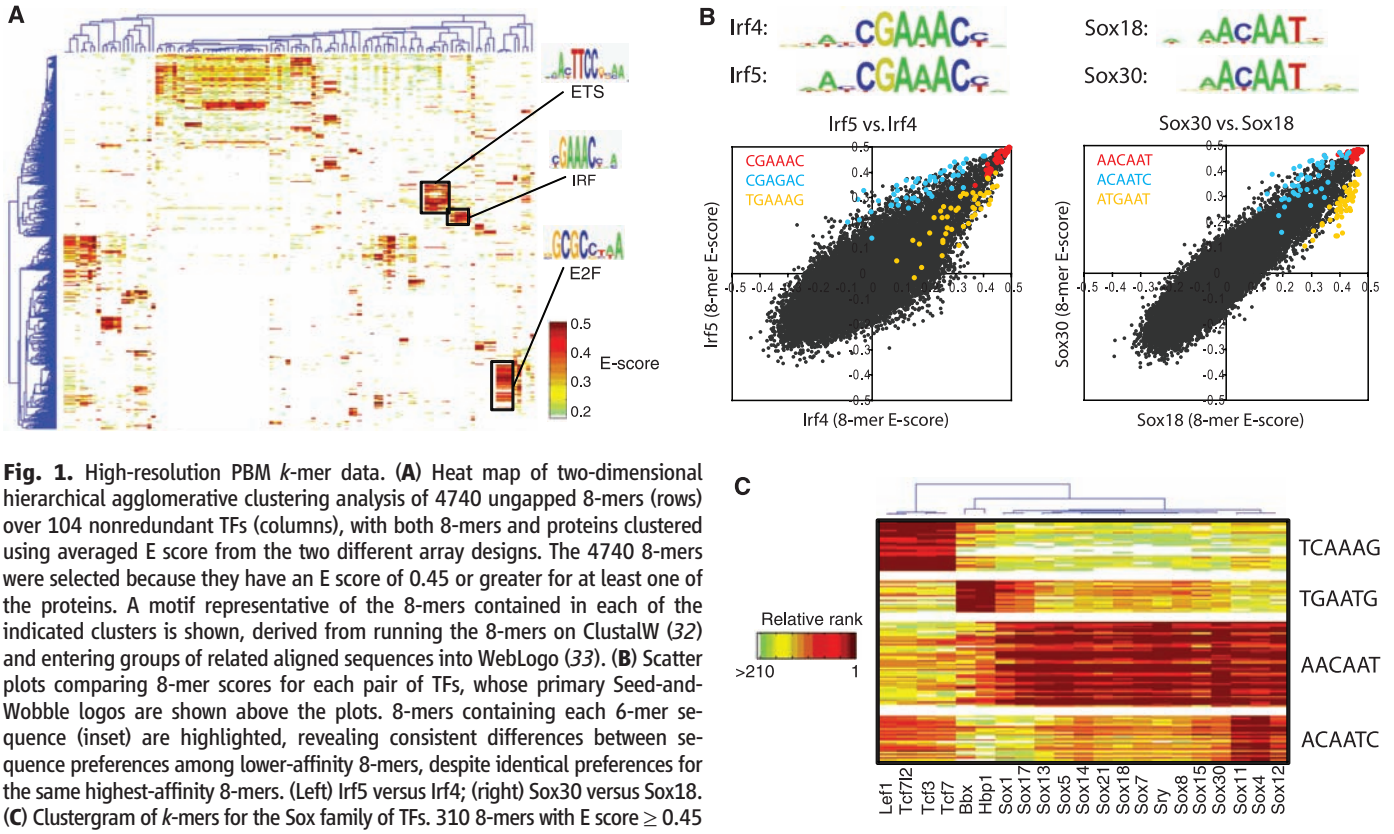

Fig. 1. High-resolution PBM *k*-mer data. (**A**) Heat map of two-dimensional hierarchical agglomerative clustering analysis of 4740 ungapped 8-mers (rows) over 104 nonredundant TFs (columns), with both 8-mers and proteins clustered using averaged E score from the two different array designs. The 4740 8-mers were selected because they have an E score of 0.45 or greater for at least one of the proteins. A motif representative of the 8-mers contained in each of the indicated clusters is shown, derived from running the 8-mers on ClustalW (*32*) and entering groups of related aligned sequences into WebLogo (*33*). (**B**) Scatter plots comparing 8-mer scores for each pair of TFs, whose primary Seed-and-Wobble logos are shown above the plots. 8-mers containing each 6-mer sequence (inset) are highlighted, revealing consistent differences between sequence preferences among lower-affinity 8-mers, despite identical preferences for the same highest-affinity 8-mers. (Left) Irf5 versus Irf4; (right) Sox30 versus Sox18. (**C**) Clustergram of *k*-mers for the Sox family of TFs. 310 8-mers with E score $\geq$ 0.45 for at least one of the 21 Sox and Sox-related TFs were hierarchically clustered according to their relative ranks for each TF, and then the rows, corresponding to *k*-mers, were rearranged to group together 8-mers with shared sequence patterns.

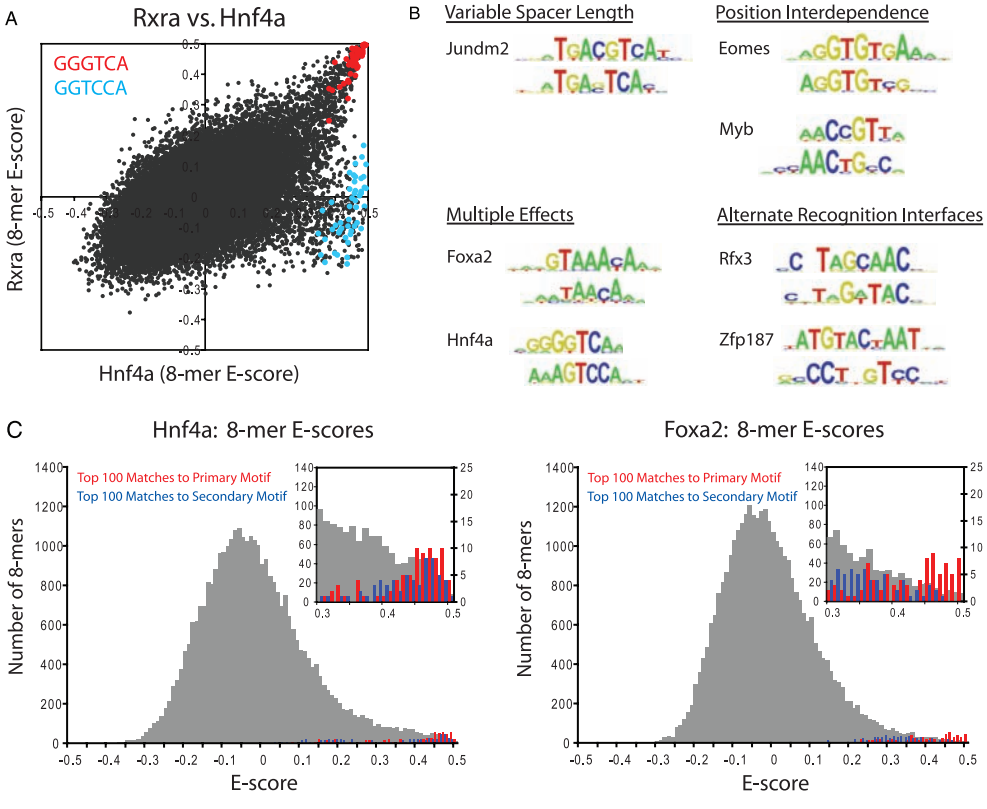

Fig. 2. TF binding site secondary motifs. (**A**) Scatter plot comparing 8-mer E scores for closely related TFs. Hnf4a and Rxra (two C4 zinc finger DBDs) both exhibit strong binding to 8-mers containing GGGTCA (red), whereas Hnf4a shows specific binding to an additional set of 8-mers containing GGTCCA (blue). (**B**) Examples of motifs from different categories of secondary motifs. (**C**) Histograms of E scores for all 8-mers (gray), the top 100 8-mer matches to the primary motif (red), and the top 100 8-mer matches to the secondary motif (blue). 8-mers were scored for matches to PWMs according to the GOMER (*27*) scoring framework. Insets provide a magnified display of the tails of the distributions; *y*-axis labels along the right of each inset refer to the red and blue bars. On the basis of the 8-mer scores, the primary and secondary Hnf4a motifs are essentially interchangeable (left), whereas Foxa2 shows a clear preference for 8-mers corresponding to its primary motif (right).

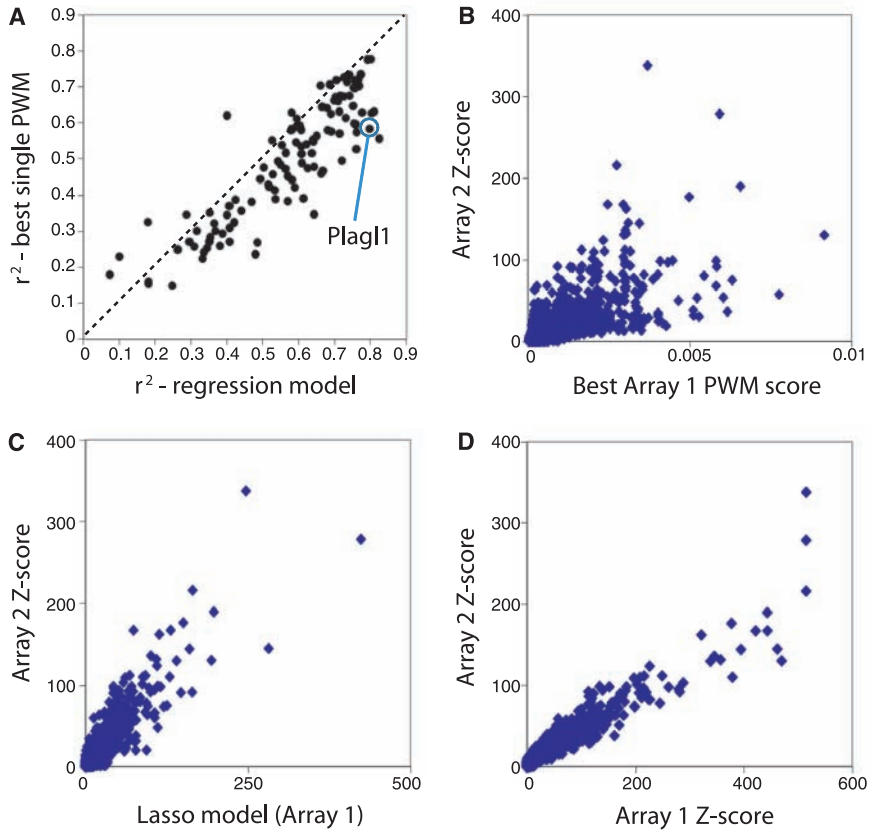

Fig. 3. Multiple-motif models typically better represent the binding profiles than do single-motif models. (**A**) Considering all TFs in this study, in general, multiple-motif models are a better representation of the data than are single-motif models. Variance in 8-mer median intensity (Z score) on Array 2 explained by our PWM regression model (*x* axis) compared to GOMER (*27*) scores for the single best PWM model obtained (best is defined as highest variance explained) over all 8-mers, with models derived from Array 1; the GOMER scoring framework calculates binding probabilities over the 8-mers according to PWMs (*27*). Each point represents one of the TFs analyzed. (**B**) The GOMER score for the best PWM derived from Array 1 is compared to the Z scores from Array 2, for Plagl1 as a case example. Each point is a single 8-mer; all 32,896 8-mers are shown. (**C**) Same as (B), except that the Array 1 regression model scores [which are a linear combination, built by using the least absolute shrinkage and selection operator (Lasso) algorithm (*34*), of GOMER scores from individual motifs] are compared to the Z scores from Array 2. (**D**) 8-mer Z scores for Plagl1 derived from Array 1 compared to the Z scores from Array 2. Each point is a single 8-mer; all 32,896 8-mers are shown.

hypothesized that the existence of secondary motifs may be a general phenomenon, and therefore, we searched for alternate binding preferences throughout our entire data set.

To aid in the identification of secondary binding preferences, we further developed our Seed-and-Wobble algorithm to search specifically for motifs that represent the *k*-mers of high signal intensity that are not explained well by the primary motif; we refer to these as the secondary motifs. A further iteration can be employed to search for a tertiary motif. As an initial test case, we examined PBM data for the human TF Oct-1 (*3*); the PBM-derived Oct-1 primary motif corresponded to the full Oct-1 DNA binding site motif, whereas the secondary and tertiary motifs corresponded to the binding site motifs of the POU_{HD} and POU_S domains (*19*), respectively (fig. S11). Analysis of 100 simulated long, 14–base pair (bp) motifs (*5*) indicated that Seed-and-Wobble was highly successful in identifying the simulated motifs and that essentially all of the secondary motifs we found in analyzing the real PBM data were unlikely to be attributable to a motif-finding artifact due to long motifs (*5*).

We observed clear secondary DNA binding preferences for nearly half of our 104 mouse TFs. Their secondary motifs fell into four different categories (Fig. 2B and supporting online material text), which we annotated manually. We confirmed binding to the secondary motifs by 6 TFs—Hnf4a, Nkx3.1, Myb, Mybl1, Foxj3, and Rfxdc2—by EMSAs (fig. S10).

We found 19 clear cases of "position interdependence" TFs, which exhibited strong interdependence (*20*) among the nucleotide positions of their binding sites. Position interdependencies frequently spanned more than just dinucleotides; for example, estrogen related receptor alpha has a strong preference for binding either CAAGGTCA or AGGGGTCA, but not CAGGGTCA or CGGGGTCA. Interdependent nucleotide positions were not always adjacent to each other; for example, Myb (fig. S10) exhibited strong interdependence at positions separated by 1 nucleotide,

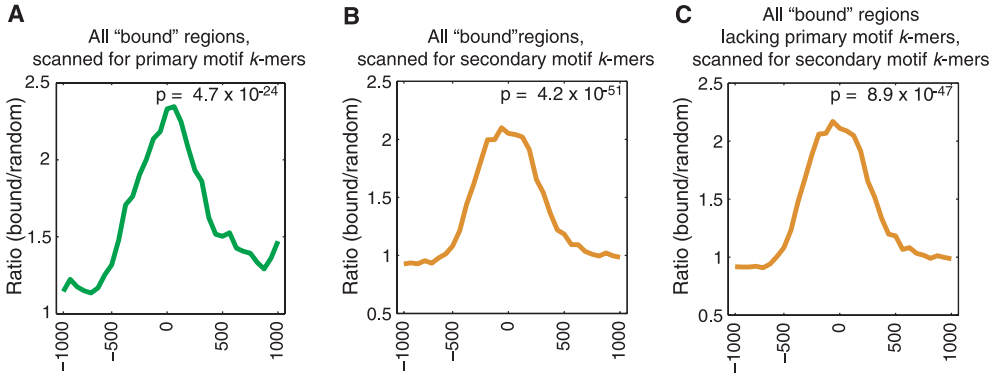

Fig. 4. Enrichment of primary versus secondary motif sequences bound in vitro within genomic regions bound in vivo. Relative enrichment of *k*-mers corresponding to the primary versus secondary Seed-and-Wobble motifs within (**A**) and (**B**) all bound genomic regions in ChIP-chip data or (**C**) those bound regions lacking primary motif *k*-mers, as compared to randomly selected sequences, was calculated (*5*) for Hnf4a (Gene Expression Omnibus accession number GSE7745). ChIP-chip "bound" peaks were identified according to the criteria of that study (*28*). A window size of 500 bp with a step size of 100 bp was used. The GOMER thresholds used are 2.958×10^{-7} and 8.419×10^{-7}, corresponding to 9 primary and 20 secondary 8-mers scanned, respectively, for Hnf4a. *P* values for enrichment of 8-mers within the bound genomic regions shown in each panel were calculated for the interval from −250 to +250 by the Wilcoxon-Mann-Whitney rank sum test, comparing the number of occurrences per sequence in the bound set versus the background set.

with preference for binding either AACCGTCA or AACTGCCA. Although position interdependence has been observed (*21*–*25*), that this phenomenon occurs on such a broad scale was not known and has important implications because commonly used TF binding site models assume mononucleotide independence.

One protein, the mouse transcriptional regulator Jundm2, which is a member of the basic leucine zipper structural class, bound to a "variable spacer length" motif (fig. S12). "Multiple effects" motifs appeared to display a combination of position interdependence and variable distances separating different parts of their motifs; at least 16 TFs fell into this category.

Finally, at least five secondary motifs in the "alternate recognition interfaces" category were not readily explainable by either a variable spacer length or position interdependence. This category is the most intriguing, as it suggests that some TFs recognize their DNA binding sites through multiple, completely different interaction modes, either through alternate structural features or by switching between alternate conformations. Support for this hypothesis comes from the co-crystal structure of human Rfx1 bound to DNA, which indicated that Rfx1 uses β strands and a connecting loop to interact with the major groove of one half-site and an α helix to interact with the minor groove of the other half-site (*26*). It is likely that Rfx3, Rfx4, and Rfxdc2 use this same mechanism of alternative DNA recognition modes (fig. S13).

For several TFs, the secondary motifs were bound nearly as well as the primary motifs, whereas in most cases, the motifs represented different affinity classes. For example, the top 20 8-mers that matched Hnf4a's primary motif were fairly evenly intermingled [$P = 0.037$ by Wilcoxon-Mann-Whitney U test, using GOMER (generalizable occupancy model of expression regulation) (*27*) scoring of motifs] with those that matched its secondary motif (Fig. 2C, left). In contrast, for Foxa2, the secondary motif represented lower-affinity binding sequences ($P = 1.94 \times 10^{-6}$) (Fig. 2C, right).

We further considered the possibility that some proteins' DNA binding specificities might be represented best by multiple motifs. We applied a linear regression approach (*5*) to learn weighted combinations of position weight matrices (PWMs) generated from several different motif-finding algorithms. We found that the binding profiles for all but 15 proteins were represented best by more than one motif (Fig. 3 and fig. S14). Some of these multiple motifs did not appear to represent different protein-DNA interaction properties described above, but nevertheless, they captured different subsets of the *k*-mer data.

We explored the in vivo usage of the secondary motifs by considering their TF occupancy. We calculated the relative enrichment of 8-mers corresponding to the primary versus secondary Seed-and-Wobble motifs within genomic regions bound in ChIP-chip data, as compared with randomly selected sequences (*5*) for Hnf4a (Fig. 4 and fig. S15, A, C, and D). As expected, Hnf4a-bound regions are enriched for matches to 8-mers corresponding to the primary motif for Hnf4a PBM data, with the greatest enrichment toward the centers of the bound regions (Fig. 4A). Hnf4a-bound regions are also enriched for matches to 8-mers corresponding to the secondary motif (Fig. 4B). Hnf4a secondary motif 8-mers are enriched even among those Hnf4a-bound regions that lack primary motif 8-mers (Fig. 4C), suggesting that the secondary motif can recruit Hnf4a to genomic loci independently of the primary motif. We observed similar results for Bcl6 (*28*) (fig. S15).

Our characterization of 104 TFs from 22 different structural classes revealed a prevalence of complexity and richness in DNA binding preferences, both across and within classes. The breadth of the observed "secondary motif" phenomenon had not been described before, and it has important implications for understanding how proteins interact with their DNA binding sites and for genome analysis.

Further experiments and analyses are needed to determine whether the same TF exerts different gene regulatory effects through distinct sequence motifs, as well as to determine whether TF-specific differences among members of a TF family (*29*) contribute to differences in binding in vivo and to distinct physiological functions. Although TFs bind a rich spectrum of *k*-mers not fully captured even by multiple PWMs, using a multiple-motif model is of practical consequence because most genome analysis tools employ PWMs. Algorithms that consider the quantitative nature of *k*-mer binding data in scoring candidate regulatory elements need to be developed.

Finally, these PBM data are likely to be highly informative for well-conserved homologs in other organisms. Generating [or inferring (*29*)] PBM data for all regulatory factors in all major model organisms is an important goal, as such *k*-mer data probably will be useful for improved prediction and analysis of regulatory elements, including the identification of direct versus indirect TF binding sites from ChIP data (*30*). Moreover, such data would aid in understanding the evolution of cis regulatory elements and transcriptional regulatory networks.

References and Notes

1. A. Tanay, *Genome Res.* **16**, 962 (2006).
2. E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, U. Gaul, *Nature* **451**, 535 (2008).
3. M. F. Berger *et al.*, *Nat. Biotechnol.* **24**, 1429 (2006).
4. M. Z. Li, S. J. Elledge, *Nat. Genet.* **37**, 311 (2005).
5. Materials and methods are available as supporting material on *Science* Online.
6. M. L. Bulyk, X. Huang, Y. Choo, G. M. Church, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 7158 (2001).
7. S. Mukherjee *et al.*, *Nat. Genet.* **36**, 1331 (2004).
8. M. F. Berger, M. L. Bulyk, *Nat. Protocols* **4**, 393 (2009).
9. X. Chen, T. R. Hughes, Q. Morris, *Bioinformatics* **23**, i72 (2007).
10. S. J. Maerkl, S. R. Quake, *Science* **315**, 233 (2007).
11. V. Matys *et al.*, *Nucleic Acids Res.* **34**, D108 (2006).
12. S. V. Orlov *et al.*, *FEBS J.* **274**, 4848 (2007).
13. J. M. Bosher, N. F. Totty, J. J. Hsuan, T. Williams, H. C. Hurst, *Oncogene* **13**, 1701 (1996).
14. S. Mertin, S. G. McDowall, V. R. Harley, *Nucleic Acids Res.* **27**, 1359 (1999).
15. M. van de Wetering, M. Oosterwegel, D. Dooijes, H. Clevers, *EMBO J.* **10**, 123 (1991).
16. A. Travis, A. Amsterdam, C. Belanger, R. Grosschedl, *Genes Dev.* **5**, 880 (1991).
17. S. G. Tevosian *et al.*, *Genes Dev.* **11**, 383 (1997).
18. K. Pfeifer, T. Prezant, L. Guarente, *Cell* **49**, 19 (1987).
19. J. D. Klemm, M. A. Rould, R. Aurora, W. Herr, C. O. Pabo, *Cell* **77**, 21 (1994).
20. P. V. Benos, M. L. Bulyk, G. D. Stormo, *Nucleic Acids Res.* **30**, 4442 (2002).
21. P. V. Benos, A. S. Lapedes, G. D. Stormo, *Bioessays* **24**, 466 (2002).
22. M. L. Bulyk, P. L. Johnson, G. M. Church, *Nucleic Acids Res.* **30**, 1255 (2002).
23. M.-L. Lee, M. Bulyk, G. Whitmore, G. Church, *Biometrics* **58**, 981 (2002).
24. T. K. Man, G. D. Stormo, *Nucleic Acids Res.* **29**, 2471 (2001).
25. Y. Barash, G. Elidan, N. Friedman, T. Kaplan, paper presented at the Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB), Berlin, 10 to 13 April 2003.
26. K. S. Gajiwala *et al.*, *Nature* **403**, 916 (2000).
27. J. A. Granek, N. D. Clarke, *Genome Biol.* **6**, R87 (2005).
28. S. M. Ranuncolo *et al.*, *Nat. Immunol.* **8**, 705 (2007).
29. M. F. Berger *et al.*, *Cell* **133**, 1266 (2008).
30. C. Zhu *et al.*, *Genome Res.* **19**, 556 (2009).
31. D. E. Newburger, M. L. Bulyk, *Nucleic Acids Res.* **37**, D77 (2009).
32. R. Chenna *et al.*, *Nucleic Acids Res.* **31**, 3497 (2003).
33. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, *Genome Res.* **14**, 1188 (2004).
34. R. Tibshirani, *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267 (1996).

35. This project was supported by funding from the Canadian Institutes of Health Research (MOP-77721 and postdoctoral fellowship to G.B.); Genome Canada through the Ontario Genomics Institute, the Ontario Research Fund, and the Canadian Institute for Advanced Research (to T.R.H.); NSF (to M.F.B.); the Canadian Foundation for Innovation and Ontario Research Fund (to Q.M.); and grant R01 HG003985 from NIH/National Human Genome Research Institute (to M.L.B.). We thank L. Peña-Castillo, A. Cheung, M. Chan, S. Bhinder, F. Bréard, P. Qureshi, S. Mnaimneh, M. Kekis, F. Khalid, J. Holroyd, D. Terterov, and K. Robasky for technical assistance and S. Gisselbrecht, K. Struhl, and S. Sunyaev for critical reading of the manuscript. PBM data are available at http://the_brain.bwh.harvard.edu/pbms/webworks/ and also via the publicly available UniPROBE database (*31*).

Supporting Online Material

www.sciencemag.org/cgi/content/full/1162327/DC1
Materials and Methods
SOM Text
Figs. S1 to S15
Tables S1 to S3
References

25 June 2008; accepted 1 May 2009
Published online 14 May 2009;
10.1126/science.1162327
Include this information when citing this paper.