Published in final edited form as:

Nat Biotechnol. 2006 November ; 24(11): 1429–1435. doi:10.1038/nbt1246.

Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities

Michael F. Berger^{1,4,7}, Anthony A. Philippakis^{1,3,4,7}, Aaron M. Qureshi^{1,5}, Fangxue S. He^{1,3}, Preston W. Estep 3rd⁶, and Martha L. Bulyk^{1,2,3,4}

¹Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115

²Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115

³Harvard/MIT Division of Health Sciences and Technology (HST), Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115

⁴Harvard University Graduate Biophysics Program, Cambridge, MA 02138

⁵Department of Mathematics, University of Maryland, College Park, MD 20742

⁶Longevity, Inc. Waltham, MA 02451

Abstract

Transcription factors (TFs) regulate the expression of genes involved in myriad cellular processes through sequence-specific interactions with DNA. In order to predict DNA regulatory elements and the TFs targeting them with greater accuracy, detailed knowledge of the binding preferences of TFs is needed. Protein binding microarray (PBM) technology permits rapid, high-throughput characterization of the *in vitro* DNA binding specificities of proteins¹. Here, we present a novel, maximally compact, synthetic DNA sequence design that represents all possible DNA sequence variants of a given length k (i.e., all “ k -mers”) on a single, universal microarray. We constructed such all k -mer microarrays covering all 10 base pair (bp) binding sites by converting high-density single-stranded oligonucleotide arrays to double-stranded DNA arrays. Using these microarrays, we comprehensively determined the binding specificities over a full range of affinities for five TFs of diverse structural classes from yeast, worm, mouse, and human. Importantly, the unbiased coverage of all k -mers permits an interrogation of binding site preferences, including nucleotide interdependencies, at unprecedented resolution.

Keywords

transcription factor; binding site motif; protein binding microarray; de Bruijn sequences; linear-feedback shift register

Correspondence should be addressed to M.L.B. (mlbulyk@receptor.med.harvard.edu).

⁷These authors contributed equally to this work.

Note: Supplementary information is available on the Nature Biotechnology website and on our website, http://the_brain.bwh.harvard.edu.

In a typical PBM experiment, a DNA-binding protein of interest is expressed with an epitope tag, purified, and applied to a double-stranded DNA microarray^{2, 3}. The microarray is then labeled with a fluorophore-conjugated antibody specific for the tag, and the binding site motif is identified from the most significantly bound spots. In a recent study, we used microarrays spotted with *Saccharomyces cerevisiae* intergenic regions (up to ~1500 bp) to identify the binding site motifs for three yeast TFs¹. The long fragments allowed a large number of sequence variants to be represented. However, a given intergenic region can contain multiple binding sites for a given TF, and we could not accurately resolve the fractional occupancies of separate sites within these regions. Moreover, yeast intergenic arrays limit the analysis to only those sequences represented in the yeast genome, and the resulting data are biased by the frequencies with which they occur. Therefore, we undertook the design of a compact, universal DNA microarray that could be used to rapidly determine the relative binding preferences of any TF from any organism for every possible binding site sequence variant.

The key to our design is two-fold. First, our double-stranded DNA probes have a length (L) considerably longer than the motif widths (k) that we intend to inspect. Thus, each spot will contain $L-k+1$ potential binding sites when considered in an overlapping fashion (Fig. 1a). Second, we have designed these spots to completely cover all possible k -mer sequence variants in a maximally compact manner, so that only a minimal number of spots needs to be synthesized. Sequences containing all 4_k overlapping k -mers exactly once are named de Bruijn sequences of order k . We utilize a special class of de Bruijn sequences generated by linear-feedback shift registers that are known to have certain advantageous pseudo-random properties⁴ (Philippakis *et al*, manuscript in preparation). We then computationally partition our de Bruijn sequence into subsequences of length L that overlap by $k-1$ to form the spots of our microarray (Fig. 1b). This construction allows substantially greater representation of sequence space than using random sequence; the same length of random sequence would miss approximately $e^{-1} \approx 37\%$ of k -mers, as dictated by the Poisson distribution. This construction also maximizes the representation of distinct sequence variants longer than k , as a de Bruijn sequence of order k will contain one-fourth of all $(k+1)$ -mers, one-sixteenth of all $(k+2)$ -mers, etc., and thus could be expected to yield substantial information about TFs with longer motif widths. At least one other group⁵ has attempted to generate sequences with maximal representation of k -mers for use in electrophoretic mobility shift assays, although their approach was limited to small k ($k \leq 5$). Another group designed an incomplete but optimized spanning set of 10-mers for PBMs to predict the binding specificity of a particular TF, but this approach required prior knowledge of the consensus binding site and was only applicable to TFs with similar binding preferences⁶. Our goal was to design a universal microarray containing all possible k -mers to investigate the binding specificities of any TF in a comprehensive, unbiased fashion.

To guide the choice of an appropriate de Bruijn sequence for our universal PBM, we first inspected the statistical properties of known TF binding sites. Examining a set of 78 TF binding site motifs from curated *in vitro* selection (SELEX) data⁷ in the JASPAR database⁸, we found that 77% had ≥ 10 informative positions (Supplementary Fig. 1). We reasoned that a de Bruijn sequence of order 10 would be suitable for the vast majority of TFs. However,

while a de Bruijn sequence of order 10 will, by definition, contain all contiguous 10-mers, it will not necessarily contain all gapped 10-mers. We observed substantial variability in the fraction of gapped k -mers represented in any single de Bruijn sequence of order k (Philippakis *et al.*, manuscript in preparation). We carefully chose our de Bruijn sequence to contain all possible 10-mers that span 11 bp with a single gap at any position (e.g., AnGGCGTTTAG, AGnGCGTTTAG, AGGnCGTTTAG, etc.). By maximizing the coverage of gapped k -mers, one simultaneously ensures that potential binding sites of widths longer than k are sampled regularly (facilitating interpolation to sites not on the array) and that the patterns containing the most informative positions in the motif are covered.

Custom ink-jet synthesized DNA microarrays containing all 10-mer binding site variants were manufactured by Agilent Technologies according to our universal design. Each microarray contains approximately 44,000 single-stranded features that are 60 nucleotides (nt) long and end-attached to the glass substrate at their 3' ends. Features were designed to begin with a single thymidine linker and a constant 24-nt primer sequence, followed by a variable 35-nt sequence. Thus, every feature contains 26 distinct, overlapping 10-mers. We note that a comparable microarray containing each 10-mer on a separate spot would require 1,048,576 probes. To prepare the microarrays for PBM experiments, we performed primer extension using unlabeled dNTPs and a small quantity of fluorescently-labeled dUTP to provide a measure of the relative amount of double-stranded DNA at each spot for subsequent data normalization (Fig. 1c). We found this process to be extremely reproducible (Supplementary Fig. 2). As a preliminary test, we performed a PBM experiment with the yeast TF Cbf1, a well-characterized regulator of the methionine biosynthesis pathway. Figure 1c shows an enlarged portion of a PBM bound by Cbf1. We note that Cbf1 possesses a basic helix-loop-helix (bHLH) DNA-binding domain and binds DNA as a homodimer, suggesting that PBMs could be used for heteromeric complexes as well as monomers.

To verify that the observed signal was due to proper Cbf1 binding, we examined all features for matches to the known 8-mer consensus site for Cbf1, GTCACGTG, and variants thereof. Since our microarray is composed of a de Bruijn sequence of order 10, every 8-mer is guaranteed to be present at least 16 times, and non-palindromic 8-mers are present at least 32 times after identifying reverse complements. As an initial approach to determining the relative binding preferences of Cbf1, we ranked all features by their normalized signal intensities and calculated the median intensity over those features with a match to each separate 8-mer. We observed that GTCACGTG exhibits the greatest median intensity of all possible 8-mers. As shown in Figure 2a, changes to certain positions within the Cbf1 binding site prove to be more detrimental to binding than others. Although any 8-mer may fall on a bright spot that also contains a true binding site, taking the median of 32 independent measurements provides an accurate estimate of the capacity of an 8-mer to be bound by a given protein.

To determine how the normalized signal intensities on our microarray correspond to TF binding affinities, we turned to the mouse TF Zif268, which contains three Cys₂His₂ zinc fingers². Zif268 has a 9-mer binding site motif, and so each possible variant is present in either orientation on at least 8 microarray features. Using various experimental methods, we and other groups have previously determined the relative binding constants for several 9-

mer DNA binding sites bound to Zif268^{2, 9, 10}. In all cases, these values are inversely correlated with the median PBM intensities, suggesting that relative signal intensities accurately estimate binding preferences (Fig. 2b, Supplementary Fig. 3).

Even though any k -mer appears on multiple spots on our microarray, these spots are variable with respect to flanking sequence and the position and orientation of the k -mer relative to the slide surface. Using control spots containing Zif268 binding sites in various positions and orientations, we observed that sites farther from the slide produced brighter signals (Supplementary Fig. 4). However, since each k -mer (for $k < 10$) represents an ensemble of measurements, we reasoned that our universal design is robust to these potential confounding variables. To test this, we designed 28 control sequences, each one containing the 9-mer Zif268 consensus site GCGTGGGCG or a single-mismatch variant, embedded in constant flanking sequence, at a fixed position and orientation relative to the slide. Each of these 28 sequences was present at 8 replicate spots, and the median of these signals was compared to the median signal of the 8 occurrences of the corresponding 9-mer in the collection of de Bruijn sequence spots (Supplementary Fig. 5). This comparison yielded a Spearman rank correlation coefficient of 0.9420, suggesting that the combined effect of flanking sequence, position, and orientation can mostly be overcome by considering multiple occurrences of the binding site.

Moreover, these contextual influences can be further minimized by performing a replicate PBM experiment on a separate microarray containing a second de Bruijn sequence. To design this second microarray, we chose a de Bruijn sequence that was uncorrelated with the first and also contained all possible 10-mers that span 11 bp with a single gap at any position. By utilizing this second array, one doubles the number of independent measurements made for each k -mer. In comparing measurements made using both arrays, we observed a striking correlation between experiments, not just for the highest-affinity binding sites, but also for moderate- and low-affinity binding sites (Fig. 2c, Supplementary Fig. 6). Further, we observed that the combined data from two independent de Bruijn sequences better captured the relative binding preferences of Zif268 than two identical PBMs of the same design (data not shown). Thus, we utilized this strategy for all TFs examined in this study.

To take advantage of this increased resolution gained from using independent de Bruijn sequences, we developed an enrichment score for individual k -mers using a modified form of the Wilcoxon-Mann-Whitney statistic (i.e., an L-statistic¹¹), as described in Supplementary Methods. Importantly, this score is: 1) easily combined for multiple PBMs of different designs, 2) robust to outliers resulting from the aforementioned position and orientation effects, and 3) invariant with respect to sample sizes, so that distinct k -mers with differing copy numbers (for example, because of differing length k) can be compared on the same scale.

To demonstrate the generality of our technology, we performed PBM experiments on a total of five TFs of diverse structural classes from different organisms: Cbf1 (bHLH from *S. cerevisiae*), Rap1 (Myb domain from *S. cerevisiae*), Ceh-22 (NK homeodomain from *Caenorhabditis elegans*), Zif268 (Cys₂His₂ zinc finger domain from mouse), and Oct-1

(POU homeodomain from human) (Fig. 3). We report all microarray probe sequences, raw and normalized signal intensities, and also our computed enrichment scores for all k -mers on our website, http://the_brain.bwh.harvard.edu. In order to compress this information on individual k -mers into a reduced motif representation, we generated a position weight matrix (PWM) for each TF via a novel, four-step method (see Supplementary Methods). An optimal 8-mer seed is identified, and the collection of all single-mismatch variants is inspected in order to identify the relative contribution of each base at each position to the binding specificity. Additional positions beyond the seed are then inspected to give a PWM (schematized in Fig. 3a). Importantly, our approach uses all of the data from the array, instead of utilizing only those features above an arbitrary cutoff, to determine the optimal motif without any prior knowledge. All DNA binding specificities determined by our universal PBMs agree well with the known sites for each TF (Fig. 3a,b). Of particular note is the yeast TF Rap1, which has been shown to recognize a motif that is approximately 12 to 13 bp in length^{1, 12, 13}. Even though our universal PBM contains only one-sixteenth of all 12-mers, we were still able to approximate the Rap1 motif using our all 10-mer microarrays.

A major assumption in the construction of a PWM is that each position within the binding site is independent¹⁴. By assessing the relative binding of any TF to every possible k -mer, our universal PBMs provide a rich dataset to test this assumption¹⁵. An advantage of our method of motif construction is that it can be adapted so that pairs of positions are varied, instead of individual positions. As an example, in our Cbf1 PBMs, GTCACTGTG (0.280) and ATCACTGTG (0.230) display the greatest enrichment scores when fixing the sequence “CACTGTG” and varying the first two positions (enrichment scores in parentheses; see Supplementary Methods); greater than GGCACGTG (0.083) and AGCACGTG (0.060). However, TTCACTGTG (−0.051) is considerably more weakly bound than TGCACGTG (0.050), suggesting an interdependence between nucleotides in the first two positions (Fig. 3c). To investigate this further, we utilized surface plasmon resonance to acquire equilibrium binding constants for these sequences. Our results confirm the observed nucleotide interdependence (Supplementary Fig. 7), suggesting that the complete binding specificity of a TF can be realized only with comprehensive measurements on all possible k -mer binding sites.

The universal “all k -mer” PBM design presented here is unique in its ability to compactly represent all k -mers and distinguish the relative binding preferences of a TF for all DNA binding site variants. The resulting data span a full range of affinities, in contrast to other techniques such as *in vitro* selection experiments⁷, from which typically only high affinity binding sites are culled. Such selection experiments have the capability to interrogate sequences of a wide range of affinities, but at the cost of increased labor and depth of sequencing. Low affinity DNA binding sites have been shown to significantly influence gene expression in several eukaryotes^{16–18}, and our data provide the opportunity to explore relationships between expression levels and binding site affinities on a genome-wide scale.

Another group recently created a microarray containing every 8-mer on self-annealing hairpinned DNAs, with one 8-mer per feature¹⁹. The far greater sequence representation afforded by our design permitted us to cover all 10-mers, and we have shown that even longer motifs can be reconstructed because of the regular sampling of longer k -mers in our

sequence design. Advances in microarray synthesis technologies that offer higher feature densities and longer feature lengths will enable comprehensive coverage of even longer binding sites; all 12-mers already could be covered by our approach using existing NimbleGen array technology²⁰ (Supplementary Fig. 8). We note that a de Bruijn sequence strategy could also be used to design RNA sequences to examine RNA binding proteins, or to design peptide sequences for peptide arrays or libraries. Additionally, our universal PBMs could be applied to the development of artificial TFs or other engineered molecules for use in therapeutics, industrial applications, or synthetic biology^{19, 21}.

Finally, multiple PBMs can be completed in parallel in a single day, providing the opportunity to obtain comprehensive TF binding site data at an unprecedented rate. We expect that as these data become available, greater insights will be gained into the function of *cis* regulatory elements and the logic of *cis* regulatory codes²².

METHODS

Protein Cloning, Expression, and Purification

All five TFs used in this study were cloned into Gateway®-compatible vectors (Invitrogen) and expressed with an N-terminal fusion to glutathione S-transferase (GST). Full-length *CBF1* was amplified from the *S. cerevisiae* genome by PCR, inserted into Gateway® donor vector pDONR201, and transferred to Gateway® destination vector pDEST15 by homologous recombination according to the manufacturer's protocols. Full-length *RAP1* was amplified from the *S. cerevisiae* genome by PCR, inserted into Gateway® donor vector pDONR221 and transferred to destination vector pDEST-GST²³. The DNA-binding domain of murine *ZIF268* (amino acids 322–435) was cloned into destination vector pDEST15-MAGIC²⁴. The DNA-binding domain of human *OCT1* (amino acids 269–440) was amplified from cDNA clone IMAGE:2966289 (accession no. BC001664) by PCR, inserted into Gateway® donor vector pENTR/d-TOPO, and transferred to destination vector pDEST15. *CEH-22* was obtained in donor vector pDONR201 from the *C. elegans* ORFeome (Open Biosystems) and transferred to destination vector pDEST15. All clones were sequence-verified to ensure that there were no mutations in the annotated DNA-binding domains.

Cbf1 and Rap1 were expressed in *Escherichia coli* strain BL21-AI (Invitrogen), and Zif268, Oct-1, and Ceh-22 were expressed in *E. coli* strain BL21-Gold(DE3)pLysS (Stratagene). Cultures were grown overnight in LB Medium containing 50 µg/ml carbenicillin, 30 µg/ml chloramphenicol (Zif268, Oct-1, Ceh-22), and 50 µM zinc acetate (Zif268), then diluted 1:100 in fresh medium. For Cbf1, cells were grown at 37°C to a final OD₂₆₀ of 0.5, induced with 0.2% L-arabinose, and incubated at 37°C for four more hours. For all other clones, cells were grown at 25°C to a final OD₂₆₀ of 0.5, induced with 1 mM isopropyl β-D-thiogalactopyranoside (IPTG), and incubated at 25°C for 14 hours. Cell pellets were collected by centrifugation at 4°C for 20 minutes at 8,000 g. Pellets were then suspended in 25 ml pre-chilled lysis buffer (Complete EDTA-free protease inhibitor tablet (Roche), 150 mM NaCl, 1 mM DTT, 50 mM Tris pH 8.0) and lysed by sonication on ice for three minutes with 30-second intervals. For Zif268, 50 µM zinc acetate was added to the lysis buffer. Cell

lysates were centrifuged once more at 4°C for 20 minutes at 30,000 g, and the soluble fractions were retained.

Proteins were purified using a 1-ml GStrap FF Column (GE Healthcare) according to the manufacturer's protocols, and eluted in 10 mM glutathione and 50 mM Tris-HCl pH 8.0. For Zif268, 50 μ M zinc acetate was added to the binding and elution buffers. Elutions of 500 μ l were collected, pooled, and concentrated approximately 10-fold using Microcon YM-30 spin columns (Millipore). The molarities of all purified proteins were determined by Western blot using a dilution series of recombinant GST (Sigma) as described in Supplementary Methods. Purified proteins were stored at -80°C until further use.

Microarray Design and Primer Extension

Custom-designed microarrays of single-stranded 60-nt oligonucleotides attached to the glass slide at the 3' ends were manufactured by Agilent Technologies. We designed the custom 44K microarrays such that nearly all 42,034 user-defined features began with a single thymidine linker attached to the slide, immediately followed by a common 24-nt sequence (3'-gtcgtgcctgtgtgcctgtgtctg-5') complementary to a common primer (5'-cagcaggacaacggaacacagac-3'). The remaining 35 nucleotides were variable sequence that contained the universal "all *k*-mer" representation. In this study, the two de Bruijn sequences utilized were generated by linear-feedback shift registers corresponding to the primitive polynomials: (1) $x^{20} + x^{19} + x^{18} + x^{16} + x^{15} + x^{13} + x^{11} + x^{10} + x^8 + x^6 + x^5 + x^4 + x^2 + x + 1$; and (2) $x^{20} + x^{19} + x^{17} + x^{15} + x^{14} + x^{12} + x^{10} + x^8 + x^7 + x^5 + x^3 + x + 1$. These polynomials were selected to generate de Bruijn sequences that not only represent all contiguous 10-mers, but also contain all 10-mers with a gap size of 1 (see Supplementary Methods). After generating each of these de Bruijn sequences of order 10 *in silico*, they were partitioned into sub-sequences corresponding to the features of the microarray of length 35 bases and overlapping by 9 bases in order to preserve all 10-mer binding sites. Thus, each feature contains 26 overlapping 10-mers, with an entire "all 10-mer" de Bruijn sequence (4^{10} binding sites) occupying 40,330 features. All microarray DNA sequences used in this study are listed on our website, http://the_brain.bwh.harvard.edu.

For primer extension of the single-stranded oligonucleotide arrays, the following were combined in a total volume of 900 μ l and heated to 85°C for 10 min: 1.17 μ M HPLC-purified common primer (Integrated DNA Technologies), 40 μ M dATP, dCTP, dGTP, and dTTP (GE Healthcare), 1.6 μ M Cy3 dUTP (GE Healthcare), 40 Units Thermo SequenaseTM DNA Polymerase (USB), and 90 μ l 10 $\times$ reaction buffer (260 mM Tris-HCl, pH 9.5, 65 mM MgCl₂). Here, the ratio of Cy3 dUTP to unlabeled dTTP was 1:25. Occasionally 1.17 nM Cy5-labeled common primer also was added to monitor the uniformity of primer annealing. A microarray, stainless steel hybridization chamber, and gasket cover slip (Agilent Technologies) were pre-warmed to 85°C for 5 min in a stationary hybridization oven. The microarray and primer extension mixture were assembled according to the manufacturer's protocols with the exception that the 900 μ l solution was sufficient to fill the entire volume of the chamber without an air bubble. The microarray was incubated at 85°C for 10 min, then 75°C for 10 min, then 65°C for 10 min, and then 60°C for 90 min. The hybridization chamber was then disassembled in a glass staining dish in 500 ml phosphate buffered saline

(PBS) / 0.01% Triton X-100 at 37°C. The microarray was transferred to a fresh staining dish, washed for 10 min in PBS / 0.01% Triton X-100 at 37°C, washed once more for 3 min in PBS at 20°C, and spun dry by centrifugation at 40 g for 6 min to dry.

Protein Binding Microarrays (PBMs)

PBM experiments were performed essentially as described previously^{1, 3}. Briefly, double-stranded microarrays were first pre-wet in PBS / 0.01% Triton X-100 for 5 mins and blocked with PBS / 2% (w/v) nonfat dried milk (Sigma) for 1 hr. Microarrays were then washed once with PBS / 0.1% Tween-20 for 5 min and once with PBS / 0.01% Triton X-100 for 2 min. Purified TFs were diluted to a final concentration of 100 nM in a 150- μ l protein binding reaction containing PBS / 2% milk / 51.3 ng/ μ l salmon testes DNA (Sigma) / 0.2 μ g/ μ l bovine serum albumin (New England Biolabs). Preincubated protein binding mixtures were applied to the microarrays and incubated for 1 hr at 20°C. Microarrays were again washed once with PBS / 0.5% Tween-20 for 10 min, and then once with PBS / 0.01% Triton X-100 for 2 min. Alexa488-conjugated rabbit anti(GST) polyclonal antibody (Molecular Probes) was diluted to 50 μ g/ml in PBS / 2% milk and applied to the microarrays for 1 hr at 20°C. Finally, microarrays were washed once with PBS / 0.05% Tween-20 for 10 min, followed by once with PBS / 0.05% Tween-20 for 3 min, and finally once with PBS for 2 min. Washed slides were spun dry by centrifuging at 40 g for 6 min. All washes were performed in Coplin jars at 20°C on an orbital shaker at 125 rpm. All microarray incubations were performed under LifterSlip coverslips (Erie Scientific) in a humid chamber. For Zif268, 50 μ M zinc acetate was added to the protein binding mixture, antibody mixture, and all wash buffers.

Microarray Analysis and Data Normalization

All microarrays were scanned (GSI Lumonics ScanArray 5000) at three different laser power settings to best capture a broad range of signal intensities and ensure subsaturation signal intensities for all spots on the microarray. Lasers of different excitation (ex) wavelengths and various emission (em) filters were used to detect different fluorophores: 633 nm ex, 670 nm em (Cy5); 543 nm ex, 570 nm em (Cy3); and 488 nm ex, 522 nm em (Alexa488). Multiple-labeled microarrays showed no interference of any fluorophores across the channels.

Microarray TIF images were quantified using GenePix Pro version 6.0 software (Molecular Devices). Bad spots (*i.e.*, spots that had scratches, dust flecks, etc.) were flagged manually and removed from subsequent analysis. For each spot, background-subtracted median intensities were calculated using the median local background. Data from multiple scans of the same slide were combined using masliner (MicroArray LINEar Regression) software, which uses subsaturated intensities to compute a linear regression for each pair of scans and extrapolate the true signal for saturated spots²⁵.

The PBM signal intensity at each spot was normalized by the corresponding amount of double-stranded DNA (Supplementary Methods). A significant incorporation bias, dependent on the local sequence context of each adenine in the template, was observed for Cy3-modified dUTP. Therefore, Cy3 intensities of the 40,330 variable de Bruijn spots were

used to compute regression coefficients for the relative contributions of all trinucleotide combinations to the total signal. Regressing over trinucleotides gave a substantially better approximation than regressing over dinucleotides, while the addition of a fourth position contributed negligibly (Supplementary Fig. 9). Using these regression coefficients, a ratio of observed-to-expected Cy3 intensity was calculated for each sequence. The PBM signal of each spot was divided by this ratio, and all spots with observed-to-expected Cy3 values less than 0.5 were removed from further consideration.

Finally, to correct for any possible non-uniformities in hybridization, these normalized PBM intensities were adjusted according to their positions on the microarray. Each spot was considered to be at the center of a block of spots seven columns wide and thirteen rows tall. (For spots closer to the margins of the microarray, the 7 x 13 block at the edge of the grid was considered.) The difference between the median normalized intensity of the spots within the block and the median normalized intensity of all spots on the microarray was subtracted from the normalized intensity at that particular spot.

Sequence Analysis and Motif Construction

Complete descriptions of methods for sequence analysis and motif construction are given in the Supplementary Methods. Briefly, for each 8-mer (either contiguous or containing up to three gapped positions), we consider the collection of all features where it occurs as a “foreground” feature set and the remaining features as a “background” feature set. We then compare foreground and background features by considering the top half (i.e., the half with highest signal intensities) from each and computing a modified form of the Wilcoxon-Mann-Whitney (WMW) statistic scaled to be invariant of foreground and background sample sizes; these two steps were necessary in order to make the statistic robust to outliers and slight changes in the number of foreground features containing a given 8-mer. We then identify the highest-scoring 8-mer with respect to this enrichment statistic, which we refer to as the “seed” of the motif. Next, at each position within this (possibly non-contiguous) 8-mer, we examined each of the four nucleotides and determined the relative contribution of each to the motif, again using a modified WMW statistic. Third, we identified the position of the 8-mer that was most degenerate, treated it as a gapped position, and extended the motif to those positions outside of the corresponding 8-mer seed. Finally, we transformed the motif derived from this method into a PWM. We note that this approach has two key advantages: 1) it utilizes information from all features in constructing the motif, instead of only choosing some fraction of the highest signal-intensity spots and weighting them equally; 2) it is able to systematically combine measurements made from multiple arrays containing different de Bruijn sequences.

Surface Plasmon Resonance

Oligonucleotide templates 60-nt in length were designed to contain one of six variant Cbf1 binding sites surrounded by degenerate flanking sequence as well as a 20-nt common primer sequence at the 3' end. All oligonucleotide sequences are listed in Supplementary Methods. Oligonucleotides were double-stranded by primer extension in the following reaction mixture: 2 μ M 60-nt oligonucleotide template (Integrated DNA Technologies), 2 μ M 20-nt 5'-biotinylated primer (Integrated DNA Technologies), 100 μ M each of dATP, dCTP,

dGTP, dTTP (GE Healthcare), 10 mM KCl, 10 mM (NH₄)₂SO₄, 20 mM Tris-HCl (pH 8.8), 2 mM MgSO₄, and 0.1% Triton X-100. Reaction mixtures were heated to 95°C for 3 min and then cooled to 60°C at 0.1 degrees/sec. Once at 60°C, 8 Units of *Bst* DNA Polymerase Large Fragment (New England Biolabs) were added, and reactions were incubated for 90 min. The resulting 60-bp products were purified by MinElute PCR Purification Kit (Qiagen). Concentrations and purity were determined by OD₂₆₀ measurements and gel electrophoresis.

Kinetic and affinity constants for TF-DNA interactions were measured using a Biacore 3000 system. Each double-stranded, biotinylated DNA sequence was immobilized to one flow cell of a streptavidin-derivatized Sensor Chip SA (Biacore) in filtered, degassed HBS-P buffer (0.01 M HEPES (pH 7.4), 0.15 M NaCl, 0.005% Surfactant P20) at 10 µl/min. Each flow cell was conjugated with approximately 50 to 60 response units of DNA. To measure the basal response to protein injection, a reference flow cell was conjugated with 60-bp DNA identical to that described above except with the Cbf1 binding site replaced with degenerate sequence. Purified GST-Cbf1 was diluted in HBS-P buffer to the following homodimer concentrations: 0.5 nM, 1 nM, 2 nM, 4 nM, 8 nM, 16 nM, 32 nM, and 64 nM. After a series of injections of HBS-P buffer, GST-Cbf1 was passed through all four flow cells of a Sensor Chip at a flow rate of 10 µl/min for 1200 seconds, followed by dissociation in empty buffer at 10 µl/min for 120 seconds. Protein samples were injected in order of increasing concentration. Between all samples, the Sensor Chip surface was regenerated by running 1 M NaCl through the flow cells for 30 seconds.

Real-time binding curves were analyzed using SCRUBBER-2 software (www.cores.utah.edu/Interaction/scrubber.html)²⁶. The binding reaction was fit to a simple bimolecular interaction model between the Cbf1 homodimer and its DNA binding site. First, the response of the reference flow cell was subtracted from each of the query flow cells, and binding curves for different concentrations were aligned to have the same injection start and stop. Binding curves then were set to zero by subtracting the response to empty buffer. For each flow cell, all curves were fit globally to a bimolecular interaction model to simultaneously determine the free parameters k_a , k_d , and R_{max} . The equilibrium binding constant (K_d) was calculated as the ratio k_d/k_a .

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank T.V.S. Murthy, Leo Brizuela, and Josh LaBaer for providing the Cbf1 and Rap1 clones, Gwenael Badis-Breard and Tim Hughes for providing the Zif268 DNA binding domain clone, and Shufen Meng for assistance with the Biacore technology. We also thank Stephen Gisselbrecht, Amy Donner, and Rachel McCord for critical reading of the manuscript. This work was funded in part by grants R01 HG003985 and R01 HG003420 from NIH/NHGRI to M.L.B. M.F.B. was supported in part by a National Science Foundation Graduate Research Fellowship. A.A.P. was supported in part by a National Defense Science and Engineering Graduate Fellowship, a National Science Foundation Graduate Research Fellowship, and an Athinoula Martinos Fellowship.

References

1. Mukherjee S, et al. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet.* 2004; 36:1331–1339. [PubMed: 15543148]
2. Bulyk ML, Huang X, Choo Y, Church GM. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci U S A.* 2001; 98:7158–7163. [PubMed: 11404456]
3. Berger MF, Bulyk ML. Protein Binding Microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA-binding proteins. *Methods in Molecular Biology.* 2006; 338:245–260. [PubMed: 16888363]
4. Golomb, S. Shift Register Sequences. Aegean Park Press; Laguna Hills, CA: 1967.
5. Kwan AH, Czolij R, Mackay JP, Crossley M. Pentaprobe: a comprehensive sequence for the one-step detection of DNA-binding activities. *Nucleic Acids Res.* 2003; 31:e124. [PubMed: 14530457]
6. Linnell J, et al. Quantitative high-throughput analysis of transcription factor binding specificities. *Nucleic Acids Res.* 2004; 32:e44. [PubMed: 14990752]
7. Oliphant AR, Brandl CJ, Struhl K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol Cell Biol.* 1989; 9:2944–2949. [PubMed: 2674675]
8. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 2004; 32:D91–94. [PubMed: 14681366]
9. Liu J, Stormo GD. Quantitative analysis of EGR proteins binding to DNA: assessing additivity in both the binding site and the protein. *BMC Bioinformatics.* 2005; 6:176. [PubMed: 16014175]
10. Miller JC, Pabo CO. Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. *J Mol Biol.* 2001; 313:309–315. [PubMed: 11800559]
11. Bjerre S. Error Bounds for Linear Combinations of Order Statistics. *Annals of Statistics.* 1977; 5:357–369.
12. Lieb JD, Liu X, Botstein D, Brown PO. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet.* 2001; 28:327–334. [PubMed: 11455386]
13. Harbison CT, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature.* 2004; 431:99–104. [PubMed: 15343339]
14. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol.* 1987; 193:723–750. [PubMed: 3612791]
15. Bulyk ML, Johnson PL, Church GM. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* 2002; 30:1255–1261. [PubMed: 11861919]
16. Jiang J, Levine M. Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell.* 1993; 72:741–752. [PubMed: 8453668]
17. Gaudet J, Mango SE. Regulation of organogenesis by the *Caenorhabditis elegans* FoxA protein PHA-4. *Science.* 2002; 295:821–825. [PubMed: 11823633]
18. Tanay A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Research.* 2006; 16:962–972. [PubMed: 16809671]
19. Warren CL, et al. Defining the sequence-recognition profile of DNA-binding molecules. *Proc Natl Acad Sci U S A.* 2006; 103:867–872. [PubMed: 16418267]
20. Singh-Gasson S, et al. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat Biotechnol.* 1999; 17:974–978. [PubMed: 10504697]
21. Blancafort P, Segal DJ, Barbas CF 3rd. Designing transcription factor architectures for drug discovery. *Mol Pharmacol.* 2004; 66:1361–1371. [PubMed: 15340042]
22. Philippakis AA, et al. Expression-guided in silico evaluation of candidate cis regulatory codes for *Drosophila* muscle founder cells. *PLoS Comput Biol.* 2006; 2:e53. [PubMed: 16733548]
23. Braun P, et al. Proteome-scale purification of human proteins from bacteria. *Proc Natl Acad Sci U S A.* 2002; 99:2654–2659. [PubMed: 11880620]

24. Li MZ, Elledge SJ. MAGIC, an in vivo genetic method for the rapid construction of recombinant DNA molecules. *Nat Genet.* 2005; 37:311–319. [PubMed: 15731760]
25. Dudley AM, Aach J, Steffen MA, Church GM. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc Natl Acad Sci U S A.* 2002; 99:7554–7559. [PubMed: 12032321]
26. Morton TA, Myszkowski DG. Kinetic analysis of macromolecular interactions using surface plasmon resonance biosensors. *Methods Enzymol.* 1998; 295:268–294. [PubMed: 9750223]
27. Wilmen A, Pick H, Niedenthal RK, Sen-Gupta M, Hegemann JH. The yeast centromere CDE1/Cpf1 complex: differences between in vitro binding and in vivo function. *Nucleic Acids Res.* 1994; 22:2791–2800. [PubMed: 8052535]
28. Christy B, Nathans D. DNA binding site of the growth factor-inducible protein Zif268. *Proc Natl Acad Sci U S A.* 1989; 86:8737–8741. [PubMed: 2510170]
29. Okkema PG, Fire A. The *Caenorhabditis elegans* NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. *Development.* 1994; 120:2175–2186. [PubMed: 7925019]
30. Klemm JD, Rould MA, Aurora R, Herr W, Pabo CO. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell.* 1994; 77:21–32. [PubMed: 8156594]

Figure 1.

Design of a universal microarray for PBM experiments. **(a)** Overlapping k -mers. Each sequence on the microarray contains several distinct, overlapping k -mer binding sites. Here, $k = 10$. **(b)** Example of a de Bruijn sequence of order 3. A de Bruijn sequence of order 3 contains all 64 3-mer variants exactly once. The de Bruijn sequence is partitioned into subsequences that overlap by 2 bases, preserving all 3-mers in the sequence. These subsequences then become the spots on the microarray. **(c)** Universal PBM containing all possible 10-mer binding sites, bound by the *S. cerevisiae* TF Cbf1 expressed with a glutathione S-transferase (GST) epitope tag. At top is a schematic showing the three main stages of each experiment: primer annealing, primer extension, and protein binding. Beneath are zoom-in images of each stage for the same microarray, scanned at different wavelengths: Cy5-labeled universal primer, Cy3-labeled dUTP, and Alexa488-conjugated α -GST antibody. Fluorescence intensities are shown in false color, with blue indicating low signal intensity, green indicating moderate signal intensity, yellow indicating high signal intensity, and white indicating saturated signal intensity. The variability observed in the Cy3-dUTP signal is due to differences in the nucleotide composition of each feature. The blank spots are single-stranded negative control probes that do not contain the universal primer sequence.

Figure 2.

Relating PBM signal intensity to individual k -mers. **(a)** Enrichment of different Cbfl binding site variants. All spots are ranked in descending order by their normalized signal intensities, and spots containing a match to each specified 8-mer are marked. For each 8-mer, the median intensity over all such spots is shown (in fluorescence units), as is the P value for enrichment as calculated by the Wilcoxon-Mann-Whitney test. **(b)** Correspondence between signal intensity and binding affinity. The median intensities for six 9-mer binding site variants for the mouse TF Zif268 are plotted against their relative dissociation constants as measured by a quantitative binding (QuMFRA) assay⁹. Data points are fit as described previously², with the addition of a constant term for nonspecific binding. **(c)** Correspondence between separate PBM experiments performed on microarrays constructed with independent de Bruijn sequences. The median intensity for spots containing a match to each 8-mer is shown for each experiment. As evident here, the PBM data are consistent not only for the highest-affinity k -mers but also for the moderate- and low-affinity k -mers. The observed correlation for 8-mers ($R^2 = 0.803$) is only slightly weaker than for 7-mers ($R^2 = 0.890$; Supplementary Fig. 6) yet considerably stronger than for 9-mers ($R^2 = 0.525$). Each non-palindromic 8-mer is present on at least 32 spots, compared to 128 and 8 spots for 7-mers and 9-mers, respectively. Differences in the absolute scales reflect differences in scanning intensities. The highest-affinity k -mers are labeled and manually aligned (inset).

Figure 3.

Determination of motifs and logos for five TFs. **(a)** Method of constructing PWMs and sequence logos, using Cbf1 as an example. First, all 8-mers containing up to three gapped positions are evaluated using our enrichment score (see **Methods**), and the highest-scoring 8-mer (in this case GTCACGTG) is used as a seed for constructing the motif. Second, at each position within this 8-mer seed, all four possible nucleotides are compared by inspecting the ranks of the probes matching each of the four variants. This analysis produces a score between -0.5 and 0.5 for each variant at each position. Third, positions outside the 8-mer seed are inspected by dropping the least informative position within the seed and repeating the preceding analysis at every additional position that yields an 8-mer with at most three gaps (ensuring that the positions inspected outside of the 8-mer seed are based on a roughly equal number of samples to those within the 8-mer seed). This analysis produces the bar graph shown. Finally, these values are converted into a sequence logo by utilizing a suitably scaled Boltzmann distribution (see Supplementary Methods). **(b)** Logos for four additional TFs constructed using this method. For each, the organism and structural class are given. Consensus sequences in panels (a) and (b) were obtained from the literature for Cbf1 (ref. 27), Zif268 (ref. 28), Ceh-22 (ref. 29), Oct-1 (ref. 30), and Rap1 (ref. 12) (standard IUPAC abbreviations are used ($K=\{T,G\}$; $R=\{A,G\}$; $Y=\{C,T\}$; $N=\{A,C,G,T\}$)). **(c)** Extension of the method for motif construction described in panel (a) to the case of di-

nucleotide variants and applied to the first two positions in the Cbf1 motif. Here, all 16 variants of the form NNCACGTG were obtained and the enrichment score of each was computed.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript