nature biotechnology

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

Michael F Berger[1,2,7], Anthony A Philippakis[1–3,7], Aaron M Qureshi[1,4], Fangxue S He[1,3], Preston W Estep III[5] & Martha L Bulyk[1–3,6]

Transcription factors (TFs) interact with specific DNA regulatory sequences to control gene expression throughout myriad cellular processes. However, the DNA binding specificities of only a small fraction of TFs are sufficiently characterized to predict the sequences that they can and cannot bind. We present a maximally compact, synthetic DNA sequence design for protein binding microarray (PBM) experiments[1] that represents all possible DNA sequence variants of a given length *k* (that is, 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 (ds) DNA arrays. Using these microarrays we comprehensively determined the binding specificities over a full range of affinities for five TFs of different structural classes from yeast, worm, mouse and human. The unbiased coverage of all *k*-mers permits high-throughput interrogation of binding site preferences, including nucleotide interdependencies, at unprecedented resolution.

In a typical PBM experiment a DNA-binding protein of interest is expressed with an epitope tag, purified, and applied to a dsDNA 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. We recently used microarrays spotted with *Saccharomyces cerevisiae* intergenic regions (up to ~1,500 bp) to identify the binding site motifs for three yeast TFs[1]. 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 analysis to those sequences represented in the yeast genome, and the resulting data are biased by the frequencies with which they occur. Therefore, we designed a compact, universal DNA microarray that could be used to rapidly determine the relative binding preferences of any TF from any organism.

There are two key aspects to our design. First, our dsDNA probes have a length (L) considerably greater 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, these spots cover all possible k-mer sequence variants in a maximally compact manner, so that it is necessary to synthesize only the minimal number of spots. Sequences containing all 4^k overlapping k-mers exactly once are named de Bruijn sequences of order k. We use a special class of de Bruijn sequences generated by linear-feedback shift registers that are known to have certain advantageous pseudo-randomness properties[4]. We then computationally partition our de Bruijn sequence into sub-sequences of length L that overlap by $k - 1$ to form the spots of our microarray (**Fig. 1b**). This allows substantially greater representation of sequence space than use of 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, in that a de Bruijn sequence of order k will contain one-fourth of all $(k + 1)$-mers, one-sixteenth of all $(k + 2)$-mers, and so on, and thus could be expected to yield substantial information about TFs with longer motif widths. At least one other group[5] has attempted to generate sequences with maximal representation of k-mers for use in electrophoretic mobility shift assays, although their approach was limited to values of $k \leq 5$. Others 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 preferences[6]. 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

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Harvard University Graduate Biophysics Program, Cambridge, Massachusetts 02138, USA. [3]Harvard/MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Department of Mathematics, University of Maryland, College Park, Maryland 20742, USA. [5]Longenity, Inc., Waltham, Massachusetts 02451, USA. [6]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to M.L.B. (mlbulyk@receptor.med.harvard.edu).

Received 6 June; accepted 28 July; published online 24 September 2006; doi:10.1038/nbt1246

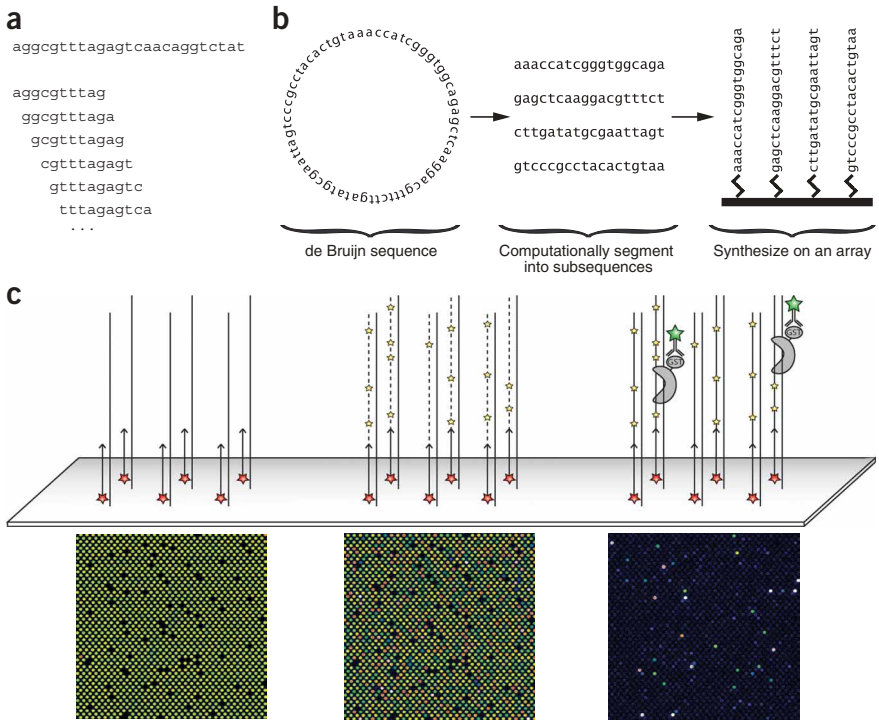

Figure 1 Design of a universal microarray for PBM experiments. (**a**) Overlapping *k*-mers. Each sequence on the microarray contains several distinct, overlapping *k*-mer binding sites. Here, $k = 10$. (**b**) Example of a de Bruijn sequence of order 3. A de Bruijn sequence of order 3 contains all 64 3-mer variants exactly once. The de Bruijn sequence is partitioned into subsequences that overlap by two 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. Above 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.

TF binding sites. Examining a set of 78 TF binding site motifs from curated *in vitro* selection (SELEX) data[7] in the JASPAR database[8], we found that 77% had ≤10 informative positions (**Supplementary Fig. 1** online). We reasoned that a de Bruijn sequence of order 10 would be suitable for the vast majority of TFs. However, although 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* (unpublished data). We chose our de Bruijn sequence to contain all possible 10-mers that span 11 bp with a single gap at any position (for example, AnGGCGTTTAG, AGnGCGTTTAG, AGGnCGTTTAG and so on). By maximizing the coverage of gapped *k*-mers, one simultaneously ensures both 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 ~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 dsDNA at each spot for subsequent data normalization (**Fig. 1c**). This process is extremely reproducible (**Supplementary Fig. 2** online). As a preliminary test, we performed a PBM experiment using the yeast TF Cbf1, a well-characterized regulator of the methionine biosynthesis pathway (**Fig. 1c**). We note that Cbf1 possesses a basic helix-loop-helix 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. Because our microarray is composed of a de Bruijn sequence of order 10, every 8-mer will occur at least 16 times, and nonpalindromic 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 shows the greatest median intensity of all possible 8-mers. Changes to certain positions within the Cbf1 binding site reduce binding more than others (**Fig. 2a**). Although any 8-mer may fall on a bright spot that also contains a true binding site, taking the median of 32 independent measurements accurately estimates the capacity of an 8-mer to be bound by a given protein.

The mouse TF Zif268, which contains three Cys_2His_2 zinc fingers[2], was used to determine how the normalized signal intensities on our microarray correspond to TF binding affinities. As Zif268 has a 9-mer binding site motif, each possible variant is present in either orientation on at least eight microarray features. Using various experimental methods, we and others have determined the relative binding constants for several 9-mer DNA-binding sites bound to Zif268 (refs. 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** online).

Even though any *k*-mer appears on multiple spots on our microarray, these spots are variable with respect to flanking sequence and

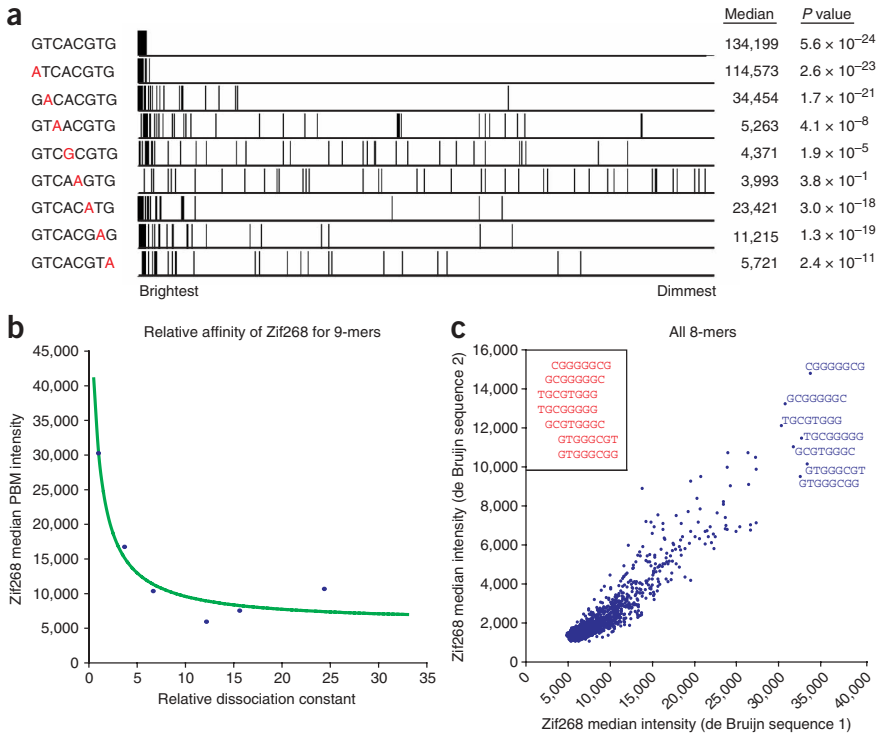

Figure 2 Relating PBM signal intensity to individual *k*-mers. (**a**) Enrichment of different Cbf1 binding site variants. All spots are ranked in descending order by their normalized signal intensities, and spots containing a match to each specified 8-mer are marked. For each 8-mer the median intensity 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) assay[9]. Data points are fitted as described previously[2], 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 *k*-mers with highest affinity but also for the *k*-mers with moderate and low affinity. 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** online) yet considerably stronger than for 9-mers (R^2 = 0.525). Each nonpalindromic 8-mer is present on at least 32 spots, compared with 128 and 8 spots for 7-mers and 9-mers, respectively. Differences in the absolute scales reflect differences in scanning intensities. The *k*-mers with highest affinity are labeled and manually aligned (inset).

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** online). However, because 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 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 occurred at eight replicate spots, and the median of these signals was compared to the median signal of the eight occurrences of the corresponding 9-mer in the collection of de Bruijn sequence spots (**Supplementary Fig. 5** online). 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 several 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 distinct de Bruijn sequence, thereby doubling the number of independent measurements made for each *k*-mer. To design this second microarray, we chose a de Bruijn sequence that was uncorrelated with the first and also contains all possible 10-mers that span 11 bp with a single gap at any position. In comparing measurements made using both arrays, we observed a striking correlation between experiments, not only for the highest affinity binding sites, but also for moderate- and low-affinity binding sites (**Fig. 2c**; **Supplementary Fig. 6** online). We also observed that the combined data from two independent de Bruijn sequences captured the relative binding preferences of Zif268 more effectively than two identical PBMs of the same design (data not shown). We therefore used this strategy for all TFs examined.

To take advantage of the 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 (that is, an *L*-statistic[11]; **Supplementary Methods** online). This score is (i) easily combined for several PBMs of different designs, (ii) robust to outliers resulting from the aforementioned position and orientation effects and (iii) 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 five TFs of diverse structural classes from different organisms: Cbf1 (basic helix-loop-helix from *S. cerevisiae*), Rap1 (Myb domain from *S. cerevisiae*), Ceh-22 (NK homeodomain from *Caenorhabditis elegans*), Zif268 (Cys_2His_2 zinc finger domain from mouse) and Oct-1 (POU homeodomain from human) (**Fig. 3**). Microarray probe sequences, raw and normalized signal intensities, and our computed enrichment scores for all *k*-mers are reported on our website (http://the_brain.bwh.harvard.edu). To compress this information on individual *k*-mers into a reduced motif representation, we generated a position weight matrix (PWM) for each TF using a four-step method (**Supplementary Methods** online). An optimal 8-mer seed is identified, and the collection of all single-mismatch variants is inspected so as 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 (**Fig. 3a**). Our approach uses all of the data from the array instead of using only those features above an arbitrary cutoff to determine the optimal motif without any prior knowledge. All DNA-binding specificities determined by our universal

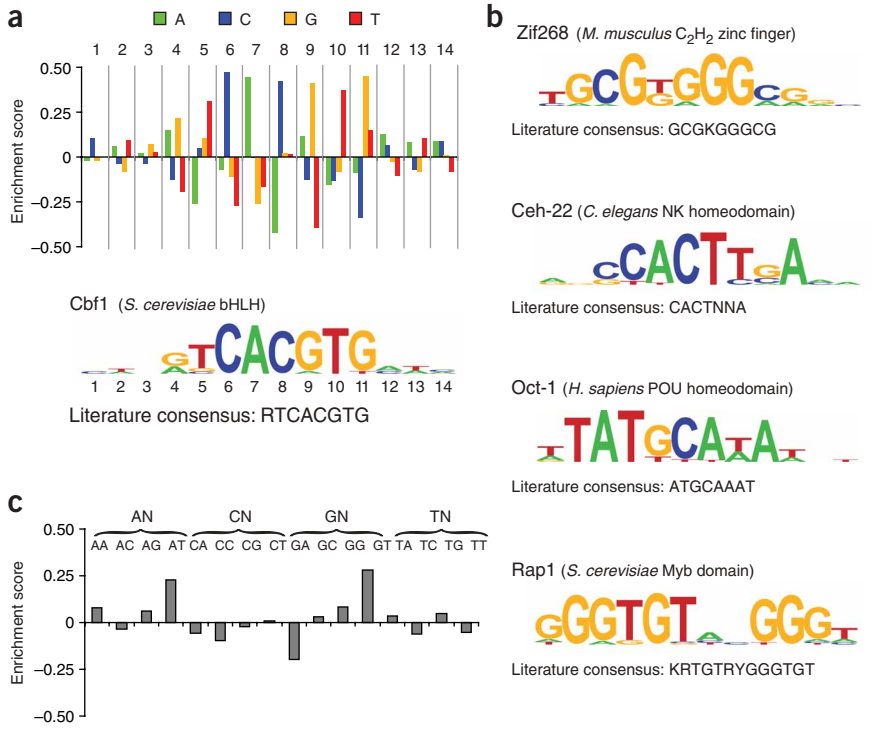

Figure 3 Determination of motifs and logos for five TFs. (**a**) Method of constructing PWMs and sequence logos, using Cbf1 as an example. First, all 8-mers containing three gapped positions or fewer are evaluated using our enrichment score (see **Methods**), and the highest scoring 8-mer (in this case 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 using a suitably scaled Boltzmann distribution (**Supplementary Methods** online). (**b**) Logos for four additional TFs constructed using this method. For each, the organism and structural class are given. Consensus sequences in **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 **a** to the case of dinucleotide 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.

PBMs agree well with the known sites for each TF (**Fig. 3a**,**b**). Of particular note is the yeast TF Rap1, which recognizes a 12- to 13-bp motif[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 independent[14]. By assessing the relative binding of any TF to every possible *k*-mer, our universal PBMs provide a rich data set to test this assumption[15]. 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, GTCACGTG (0.280) and ATCACGTG (0.230) display the greatest enrichment scores when fixing the sequence CACGTG and varying the first two positions (enrichment scores in parentheses; **Supplementary Methods** online)—greater than GGCACGTG (0.083) and AGCACGTG (0.060). However, binding of Cbf1 to TTCACGTG (–0.051) is considerably weaker than binding to TGCACGTG (0.050), suggesting an interdependence between nucleotides in the first two positions (**Fig. 3c**). To investigate this further, we used surface plasmon resonance to acquire equilibrium binding constants for these sequences. Our results confirm the observed nucleotide interdependence (**Supplementary Fig. 7** online), 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 experiments[7], 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 feature[19]. 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 permit comprehensive coverage of even longer binding sites; all 12-mers already could be covered by our approach using existing NimbleGen array technology[20] (**Supplementary Fig. 8** online). 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. 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 PBM experiments can be completed in parallel in a single day, providing the opportunity to obtain comprehensive TF binding site data at an unprecedented rate. The generation of these data will enhance insight into the function of *cis* regulatory elements and the logic of *cis* regulatory codes[22].

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 polymerase chain reaction (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-GST[23]. The DNA-binding domain of murine *Zif268* (amino acids 322–435) was cloned into destination vector pDEST15-MAGIC[24]. The DNA-binding domain of human *OCT1* (amino acids 269–440) was amplified from complementary DNA clone IMAGE:2966289 (accession number 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 Luria-Bertani 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 OD_{260} of 0.5, induced with 0.2% L-arabinose and incubated at 37 °C for an additional 4 h. For all other clones, cells were grown at 25 °C to a final OD_{260} of 0.5, induced with 1 mM isopropyl β-D-thiogalactopyranoside and incubated at 25 °C for 14 h. Cell pellets were collected by centrifugation at 4 °C for 20 min at 8,000*g*. Pellets were then suspended in 25 ml prechilled lysis buffer (Complete EDTA-free protease inhibitor tablet (Roche), 150 mM NaCl, 1 mM dithiothreitol, 50 mM Tris, pH 8.0) and lysed by sonication on ice for 3 min with 30-s intervals. For Zif268, 50 μM zinc acetate was added to the lysis buffer. Cell lysates were centrifuged once more at 4 °C for 20 min 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 μM zinc acetate was added to the binding and elution buffers. Elutions of 500 μl were collected, pooled and concentrated ~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 44,000 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′-gtcgtgcctgttgccttgtgtctg-5′) complementary to a common primer (5′-cagcacggacaacggaacacagac-3′). The remaining 35 nt were of variable sequence that contained the universal 'all *k*-mer' representation. In this study the two de Bruijn sequences used were generated by linear-feedback shift registers corresponding to the primitive polynomials (i) $x^{20} + x^{19} + x^{18} + x^{16} + x^{15} + x^{13} + x^{11} + x^{10} + x^8 + x^6 + x^5 + x^4 + x^2 + x + 1$ and (ii) $x^{20} + x^{19} + x^{17} + x^{15} + x^{14} + x^{12} + x^{10} + x^8 + x^7 + x^5 + x^3 + x + 1$. 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 so as 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. The array sequence sets used in this study are freely available at our lab website (http://the_brain.bwh.harvard.edu) under an academic license to researchers at academic or non-profit institutions. Any commercial synthesis or use of such arrays must be the subject of a separate licensing agreement with Brigham & Women's Hospital's Office of Corporate Sponsored Research and Licensing (CSRL).

For primer extension of the single-stranded oligonucleotide arrays the following were combined in a total volume of 900 μl and heated to 85 °C for 10 min: 1.17 μM high-performance liquid chromatography–purified common primer (Integrated DNA Technologies), 40 μM dATP, dCTP, dGTP and dTTP (GE Healthcare), 1.6 μM cyanine 3 (Cy3) dUTP (GE Healthcare), 40 units Thermo Sequenase DNA Polymerase (USB) and 90 μl 10× reaction buffer (260 mM Tris-HCl, pH 9.5, 65 mM $MgCl_2$). 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 coverslip (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-μl 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% (vol/vol) 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.

PBMs. PBM experiments were performed essentially as described[1,3]. Briefly, double-stranded microarrays were first premoistened in PBS–0.01% Triton X-100 for 5 min and blocked with PBS–2% (wt/vol) nonfat dried milk (Sigma) for 1 h. Microarrays were then washed once with PBS–0.1% (vol/vol) 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-μl protein binding reaction containing PBS–2% (wt/vol) milk –51.3 ng/μl salmon testes DNA (Sigma)–0.2 μg/μl bovine serum albumin (New England Biolabs). Preincubated protein binding mixtures were applied to the microarrays and incubated for 1 h at 20 °C. Microarrays were again washed once with PBS–0.5% (vol/vol) Tween-20 for 10 min, and then once with PBS–0.01% Triton X-100 for 2 min. Alexa488-conjugated rabbit polyclonal antibody to GST (Molecular Probes) was diluted to 50 μg/ml in PBS–2% milk and applied to the microarrays for 1 h at 20 °C. Finally, microarrays were washed once with PBS–0.05% (vol/vol) Tween-20 for 10 min and then 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 r.p.m. All microarray incubations were performed under LifterSlip coverslips (Erie Scientific) in a humid chamber. For Zif268, 50 μM 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 TIFF (tagged image file format) images were quantified using GenePix Pro Version 6.0 software (Molecular Devices). Bad spots (that is, spots that had scratches, dust flecks or other imperfections) 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 spots[25].

The PBM signal intensity at each spot was normalized by the corresponding amount of dsDNA (**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, whereas the addition of a fourth position contributed negligibly (**Supplementary Fig. 9** online). 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 nonuniformities 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 7 columns wide and 13 rows tall. (For spots closer to the margins of the microarray, the 7 × 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 three gapped positions or fewer) 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 (that is, the half with highest signal intensities) from each and computing a modified form of the Wilcoxon-Mann-Whitney statistic scaled to be invariant of foreground and background sample sizes; these two steps were necessary 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 noncontiguous) 8-mer, we examined each of the four nucleotides and determined the relative contribution of each to the motif, again using a modified Wilcoxon-Mann-Whitney 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. Not only does it use information from all features in constructing the motif, instead of only choosing some fraction of the spots of highest signal intensity and weighting them equally, but it is also 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 made double stranded by primer extension in the following reaction mixture: 2 μM 60-nt oligonucleotide template (Integrated DNA Technologies), 2 μM 20-nt 5′-biotinylated primer (Integrated DNA Technologies), 100 μM each of dATP, dCTP, dGTP, dTTP (GE Healthcare), 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 20 mM Tris-HCl (pH 8.8), 2 mM $MgSO_4$ and 0.1% (vol/vol) Triton X-100. Reaction mixtures were heated to 95 °C for 3 min and then cooled to 60 °C at 0.1 degree per second. Once the target temperature of 60 °C had been reached, 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 OD_{260} 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% (wt/vol) Surfactant P20) at 10 μl/min. Each flow cell was conjugated with ~50–60 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 just described 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 1,200 s and then allowed to dissociate in empty buffer at 10 μl/min for 120 s. 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 s.

Real-time binding curves were analyzed using SCRUBBER-2 software (http://www.cores.utah.edu/Interaction/scrubber.html)[26]. The binding reaction was fitted 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 normalized by subtracting the response to empty buffer. For each flow cell all curves were fitted 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.

Note: Supplementary information is available on the Nature Biotechnology website.

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 National Institutes of Health/National Human Genome Research Institute 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.

AUTHOR CONTRIBUTIONS

M.F.B. participated in the array design, experimental design, analysis of results and drafting of the manuscript, and performed the experiments; A.A.P. conceived the idea of using de Bruijn sequences and participated in the binding site survey, array design, experimental design, analysis of results and drafting of the manuscript; A.M.Q. provided linear feedback shift register expertise and assisted in the array design; F.S.H. participated in the binding site survey; P.W.E. III conceived the concept of the compact universal array; M.L.B. conceived the concept of the compact universal array and participated in the array design, experimental design, analysis of the results and drafting of the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Biotechnology* website for details).

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Mukherjee, S. *et al.* Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* **36**, 1331–1339 (2004).
2. Bulyk, M.L., Huang, X., Choo, Y. & Church, G.M. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl. Acad. Sci. USA* **98**, 7158–7163 (2001).
3. Berger, M.F. & Bulyk, M.L. Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA-binding proteins. *Methods Mol. Biol.* **338**, 245–260 (2006).
4. Golomb, S.. *Shift Register Sequences* (Aegean Park Press, Laguna Hills, California, USA, 1967).
5. Kwan, A.H., Czolij, R., Mackay, J.P. & Crossley, M. Pentaprobe: a comprehensive sequence for the one-step detection of DNA-binding activities. *Nucleic Acids Res.* **31**, e124 (2003).
6. Linnell, J. *et al.* Quantitative high-throughput analysis of transcription factor binding specificities. *Nucleic Acids Res.* **32**, e44 (2004).
7. Oliphant, A.R., Brandl, C.J. & 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.* **9**, 2944–2949 (1989).

8. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).
9. Liu, J. & Stormo, G.D. Quantitative analysis of EGR proteins binding to DNA: assessing additivity in both the binding site and the protein. *BMC Bioinformatics* **6**, 176 (2005).
10. Miller, J.C. & Pabo, C.O. Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger–DNA recognition. *J. Mol. Biol.* **313**, 309–315 (2001).
11. Bjerve, S. Error bounds for linear combinations of order statistics. *Ann. Stat.* **5**, 357–369 (1977).
12. Lieb, J.D., Liu, X., Botstein, D. & Brown, P.O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat. Genet.* **28**, 327–334 (2001).
13. Harbison, C.T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
14. Berg, O.G. & von Hippel, P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723–750 (1987).
15. Bulyk, M.L., Johnson, P.L. & Church, G.M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* **30**, 1255–1261 (2002).
16. Jiang, J. & Levine, M. Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell* **72**, 741–752 (1993).
17. Gaudet, J. & Mango, S.E. Regulation of organogenesis by the *Caenorhabditis elegans* FoxA protein PHA-4. *Science* **295**, 821–825 (2002).
18. Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.* **16**, 962–972 (2006).
19. Warren, C.L. *et al.* Defining the sequence-recognition profile of DNA-binding molecules. *Proc. Natl. Acad. Sci. USA* **103**, 867–872 (2006).
20. Singh-Gasson, S. *et al.* Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.* **17**, 974–978 (1999).
21. Blancafort, P., Segal, D.J. & Barbas, C.F., III. Designing transcription factor architectures for drug discovery. *Mol. Pharmacol.* **66**, 1361–1371 (2004).
22. Philippakis, A.A. *et al.* Expression-guided *in silico* evaluation of candidate *cis* regulatory codes for *Drosophila* muscle founder cells. *PLoS Comput. Biol.* **2**, e53 (2006).
23. Braun, P. *et al.* Proteome-scale purification of human proteins from bacteria. *Proc. Natl. Acad. Sci. USA* **99**, 2654–2659 (2002).
24. Li, M.Z. & Elledge, S.J. MAGIC, an *in vivo* genetic method for the rapid construction of recombinant DNA molecules. *Nat. Genet.* **37**, 311–319 (2005).
25. Dudley, A.M., Aach, J., Steffen, M.A. & Church, G.M. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc. Natl. Acad. Sci. USA* **99**, 7554–7559 (2002).
26. Morton, T.A. & Myszka, D.G. Kinetic analysis of macromolecular interactions using surface plasmon resonance biosensors. *Methods Enzymol.* **295**, 268–294 (1998).
27. Wilmen, A., Pick, H., Niedenthal, R.K., Sen-Gupta, M. & Hegemann, J.H. The yeast centromere CDEI/Cpf1 complex: differences between *in vitro* binding and *in vivo* function. *Nucleic Acids Res.* **22**, 2791–2800 (1994).
28. Christy, B. & Nathans, D. DNA binding site of the growth factor–inducible protein Zif268. *Proc. Natl. Acad. Sci. USA* **86**, 8737–8741 (1989).
29. Okkema, P.G. & Fire, A. The *Caenorhabditis elegans* NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. *Development* **120**, 2175–2186 (1994).
30. Klemm, J.D., Rould, M.A., Aurora, R., Herr, W. & Pabo, C.O. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell* **77**, 21–32 (1994).