

Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors

Michael F Berger^{1,2} & Martha L Bulyk¹⁻⁴

¹Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. ²Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, Massachusetts 02138, USA. ³Harvard-MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, Massachusetts 02115, USA. ⁴Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. Correspondence should be addressed to M.L.B. (mlbulyk@receptor.med.harvard.edu).

Published online 5 March 2009; doi:10.1038/nprot.2008.195

Protein-binding microarray (PBM) technology provides a rapid, high-throughput means of characterizing the *in vitro* DNA-binding specificities of transcription factors (TFs). Using high-density, custom-designed microarrays containing all 10-mer sequence variants, one can obtain comprehensive binding-site measurements for any TF, regardless of its structural class or species of origin. Here, we present a protocol for the examination and analysis of TF-binding specificities at high resolution using such 'all 10-mer' universal PBMs. This procedure involves double-stranding a commercially synthesized DNA oligonucleotide array, binding a TF directly to the double-stranded DNA microarray and labeling the protein-bound microarray with a fluorophore-conjugated antibody. We describe how to computationally extract the relative binding preferences of the examined TF for all possible contiguous and gapped 8-mers over the full range of affinities, from highest affinity sites to nonspecific sites. Multiple proteins can be tested in parallel in separate chambers on a single microarray, enabling the processing of a dozen or more TFs in a single day.

INTRODUCTION

Cells respond to environmental stimuli, progress through the cell cycle and adapt to changes in growth conditions by altering the expression of particular genes across the genome. In multicellular organisms, spatial and temporal changes in gene expression throughout development enable the formation of organs and tissues consisting of morphologically and functionally diverse cell types. Gene expression levels are dynamically regulated by TFs through sequence-specific interactions with genomic DNA. As master regulators of numerous cellular processes, TFs constitute a substantial presence in the gene complement of every organism, accounting for approximately 5–10% of genes in eukaryotes¹⁻⁵. These proteins may function as either activators or repressors and may bind alone or in combination near the genes whose expression they control. The binding sites for eukaryotic TFs are themselves typically short (6–10 base pairs) and often exhibit considerable degeneracy. To globally map TFs to their target genes and understand the regulatory interactions that govern cellular identity and behavior, precise knowledge of the full range of the DNA-binding specificities of TFs is necessary. Despite their central importance, however, comprehensive binding-site measurements have been obtained for only a small number of TFs. Existing binding data are typically sparse, with only a handful of sites having been experimentally determined for any TF, and they frequently exhibit ascertainment bias according to affinity or simply which binding sites happened to have been identified first. Consequently, predictions of regulatory elements across the genome on the basis of these limited binding data are prone to false positives and false negatives. Further, the binding specificities of the majority of eukaryotic TFs are currently completely unknown.

We have developed PBM technology as a rapid, high-throughput means of characterizing the sequence specificities of DNA–protein interactions *in vitro*⁶⁻⁹. In contrast to earlier *in vitro* technologies

for examining DNA–protein interactions (see below), which have been time consuming and not highly scalable, PBMs enable the simultaneous measurement of the relative affinities of a TF for tens of thousands of individual DNA sequences in less than a day. In a typical PBM experiment, a purified, epitope-tagged TF is allowed to bind directly to a double-stranded DNA microarray, and the protein-bound array is labeled with a fluorophore-conjugated antibody specific to the epitope, providing a quantitative readout of the relative amounts of protein bound to each of the probe sequences on the array¹⁰. Intrinsic sequence preferences for the TF can be extracted according to the enrichment of these sequences among the brightest probes on the array.

The microarrays themselves can be fabricated in various ways. Microarrays spotted with a limited number of short, double-stranded DNA oligonucleotides were used previously to monitor the relative preference of wild-type and various mutant constructs of the mouse TF Egr1 (Zif268) for 64 variant binding sites⁷. We first extended the technique to the genome scale by spotting long PCR products representing all intergenic regions of the *Saccharomyces cerevisiae* genome to map the binding sites for a number of structurally diverse TFs from yeast⁸. For TFs of other organisms, however, yeast intergenic arrays limit the analysis to only those sequences represented in the *S. cerevisiae* genome, and the resulting data are biased by the frequencies with which those sequences occur on the arrays. Moreover, a given intergenic region can contain multiple binding sites for a given TF, complicating the accurate resolution of the fractional occupancies of separate sites within the lengthy DNA fragments.

Here, we describe experimental and data analysis protocols for a universal PBM platform that uses synthetic (nongenic) sequence to achieve both the desired versatility and binding-site resolution for use in a new generation of PBM assays. We have

specially designed our universal PBMs to contain all possible 10-bp sequences in a space- and cost-efficient manner^{9,11}. As such, they can be used to comprehensively characterize the full range of binding specificity of any TF from any structural family in any species, as long as the TF is capable of binding to sites that have ~12 or fewer informative nucleotide positions. (At this time, it is uncertain whether our 'all 10-mer' PBM assays can derive the binding specificities of TFs that bind significantly longer DNA-binding site motifs.) Custom-designed microarrays are synthesized by Agilent Technologies in an array of single-stranded 60-mer probes, and they are subsequently double-stranded biochemically in a solid-phase primer extension reaction before protein binding and antibody labeling (Fig. 1). Probe signal intensities from a protein-bound microarray can be deconvoluted to produce a measure of the relative affinity of the TF for all *k*-mers (i.e., 'words,' or DNA sequences of length *k*). Currently available array formats from Agilent enable the physical separation of a single slide into multiple chambers for separate PBM experiments. Consequently, binding data can be rapidly generated for large numbers of TFs, with each individual data set depicting an extremely rich landscape of sequence preferences encompassing both high- and low-affinity sites.

By providing comprehensive measurements for all possible binding-site variants, universal PBMs offer the potential for improved computational methods of TF-binding specificity representation and binding-site discovery. Traditionally, TF-binding specificities have been represented as either International Union of Pure and Applied Chemistry consensus DNA sequences or mononucleotide position weight matrices (PWMs)¹². Both forms are typically based on a limited number of known binding sites, from which the preferences of the TF for all other sequences are approximated. Furthermore, standard mononucleotide PWMs are based on the assumption that all positions within the motif exert additive, independent effects on binding affinity. It has been shown that this is not the case for certain TFs, where the nucleotide preference at one position depends on which particular nucleotide occupies another position^{13–15}. With universal PBMs, however, the binding specificity of a TF is more accurately captured in a look-up table that conveys its relative preference for every individual 'word.' Nucleotide interdependence information is retained, and both high- and low-affinity classes of sites are identifiable. Nevertheless, we present here one approach for compactly representing PBM-binding data in a PWM that uses the unbiased sequence coverage on the array to identify the relative contribution of each nucleotide at each position to the binding specificity.

In addition to providing a biochemical representation of TF–DNA interactions *in vitro*, PBMs can provide biological insights into the *in vivo* functions and regulatory roles of TFs. Gene regulation involves the dynamic association and dissociation of TFs and their binding sites *in vivo*. Consequently, to map and fully understand the regulatory interactions that underlie the global patterns of gene expression in an organism, one would need to know which binding sites throughout the genome are used in every cellular state and environmental perturbation. Methods to directly measure genome-wide TF occupancy *in vivo* have proven very useful (see below), but they are often hindered by experimental limitations, and examining every TF under all possible cell types and/or conditions is not feasible, particularly given potentially infinite 'condition space.' Alternatively, universal PBMs enable the

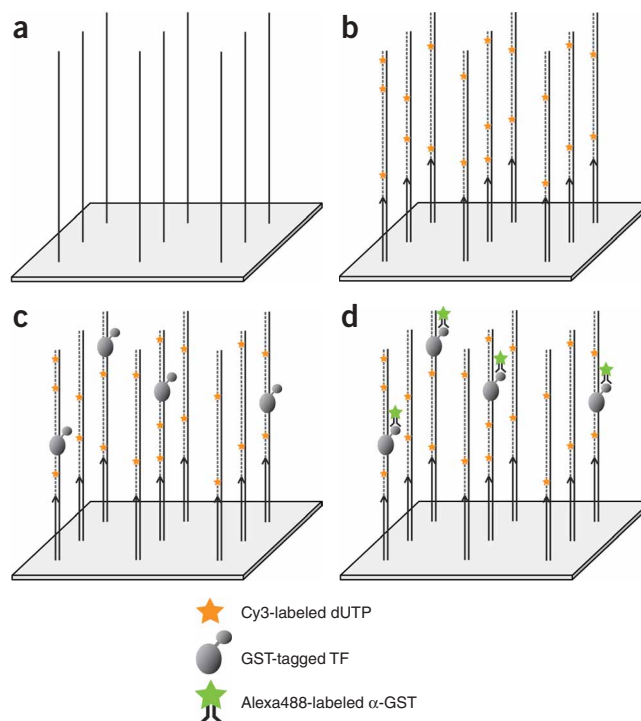


Figure 1 | Schematic of universal PBM experiments. (a) A commercially synthesized single-stranded DNA microarray is double-stranded by (b) solid-phase primer extension using a small amount of spiked-in fluorescently labeled dUTP. (c) An epitope-tagged TF is bound directly to the DNA on the microarray, and the (d) protein-bound array is labeled with a fluorophore-conjugated antibody.

rapid identification of all possible binding-site sequence variants in a single experiment. These binding data can be subsequently integrated with global gene expression profiles to infer the condition-specific targets and functions of TFs¹⁶. The *in vitro* binding specificities derived from universal PBM experiments show good agreement with preferred *in vivo* sites, when known¹⁷. Given the speed and ease with which these experiments can be performed, *in vitro* binding data can readily be generated for large numbers of TFs. This is noteworthy, considering that TFs number approximately 300 in *S. cerevisiae*¹, 750 in *Drosophila melanogaster*³ and almost 2,000 in humans⁵. Furthermore, the combinatorial nature of gene regulation in higher eukaryotes necessitates the creation of a large catalog of TF-binding sites to locate potential regulatory sequences and understand the regulatory relationships that exist.

Comparison with other methods

Several other methods exist for determining the *in vitro* DNA-binding specificities of TFs. Electrophoretic mobility shift assay^{18,19}, DNase I footprinting²⁰, southwestern blotting²¹ and surface plasmon resonance²² are predominantly low-throughput approaches for examining a small number of distinct DNA sequences and exhibit different levels of precision. *In vitro* selection²³ has been used to identify larger sets of binding sequences. This process involves an initial *in vitro* selection from a randomized pool of DNA oligonucleotides, followed by several additional cycles of amplification, selection and ultimately sequencing. Like universal PBMs, this approach can provide an unbiased collection of

permissible DNA-binding site sequences; however, in most applications, only the highest-affinity sequences are retained for sequencing. These approaches are not currently suitable for the acquisition of comprehensive binding data for all sequence variants.

Protein-binding microarray technology has been adapted by other groups on a small scale to determine the *in vitro* binding preferences of particular TFs or TF families^{24,25}. On a larger scale, Ansari and colleagues²⁶ synthesized a microarray composed of self-annealing hairpin probes covering all 8-mers (one 8-mer per probe), to which they bound small molecules as well as a TF in a PBM-like assay. These experiments provide similar information as universal PBMs, although the greater sequence coverage afforded by our compact combinatorial design permits the recovery of the DNA-binding preferences of TFs with longer and/or gapped motifs. Other microarray-based approaches have been developed to determine the biochemical affinity of a TF for its many target sequences. DNA microarrays coupled with surface plasmon resonance have been used to simultaneously monitor the kinetics of binding of the yeast TF Gal4 to 120 double-stranded DNA molecules²⁷. Maerkl and Quake²⁸ recently designed a microfluidic device that enabled them to measure the equilibrium dissociation constants of 4 TFs for 256 different DNA sequences. Both of these methods require previous knowledge of the binding specificity of a TF and the design of separate sets of probe sequences to examine different TFs or TF families due to the limited throughput of each technology. Furthermore, we have observed that universal PBM fluorescence signal intensities are generally proportional to relative affinity; however, the precise relationship between signal intensity and absolute affinity is still under investigation.

Methods to monitor the *in vivo* occupancy of TF-binding sites across the genome produce data complementary to those from PBMs. ChIP-chip^{29–31}, or chromatin immunoprecipitation coupled with microarray hybridization, provides a direct measure of *in vivo* DNA interactions in a given cell type at a given time point and has been successfully used to examine TF binding in numerous organisms for a variety of conditions and tissues³². A separate microarray-based technique, DamID, uses a fusion protein between a TF and DNA adenine methyltransferase (Dam) and relies on detection of genomic DNA after digestion with a methylation-sensitive restriction enzyme³³. ChIP-Seq³⁴ and ChIP-PET³⁵ both use high-throughput sequencing as a readout of chromatin immunoprecipitated DNA, which can facilitate the mapping of bound regions to a larger fraction of the genome and at higher resolution than contemporary microarray hybridization³⁶. These high-throughput *in vivo* approaches, although valuable, do possess certain technical limitations, such as the availability of ChIP-grade antibody and the accessibility of the epitope upon binding to DNA (for ChIP), as well as potentially limiting tissue sources. The interactions identified by these methods may not always correspond to direct protein–DNA contacts but could instead result from indirect association mediated by several intermediate proteins or complexes. Resolution is also limited due to difficulties in reducing the size of DNA fragments (ChIP-chip) or to the spread of methylation (DamID). Finally, these *in vivo* experiments must be conducted under conditions in which the TF of interest is expressed, nuclear, and actively bound to its target sites. Such conditions are not always known *a priori*, and TFs typically respond to many conditions and stimuli, such that it is impractical to examine every possible cellular state to fully map all functional

interactions. The *in vitro* nature of PBMs eliminates many of the technical limitations of *in vivo* approaches, and PBM experiments for multiple proteins can be completed rapidly in less than a day. Furthermore, we have found the binding specificities derived from universal PBM experiments to be very consistent with known *in vivo* binding sites for well-studied TFs¹⁷. Although PBMs themselves do not directly identify genomic loci bound by a TF *in vivo* in a particular cellular condition, PBMs can be used to capture all possible binding sites in a single experiment. These data can then be integrated with genomic sequence, global gene expression profiles and other data types to infer functional binding-site usage in various conditions.

Applications of universal PBMs

Given the abundance of TFs in the gene complement of every organism, universal PBMs can be used directly for the characterization of the binding specificities of thousands of individual TFs. As of this writing, universal PBMs have been used to interrogate the sequence preferences of TFs from prokaryotic and eukaryotic species, including *Vibrio harveyi*³⁷, *Plasmodium falciparum*³⁸, *S. cerevisiae*⁹, *Caenorhabditis elegans*⁹, *D. melanogaster* (M.L.B. Lab and A.M. Michelson Lab, unpublished data), mouse^{9,17,39} and humans⁹. Moreover, in addition to characterizing the DNA-binding specificities of each individual protein, PBMs can be adapted to study the DNA-binding specificities of heterodimers (F. De Masi and M.L.B., unpublished data) and the influence of ligands and protein cofactors on DNA binding⁴⁰. Alterations in the overall affinity or even intrinsic sequence preferences of a TF could be monitored in the presence and absence of ligand, in combination with multiple dimerization partners and in multiprotein complexes.

By providing comprehensive measurements for all possible *k*-mer sequence variants, universal PBMs offer the opportunity to examine the full landscape of TF binding at high resolution. Accordingly, families of TFs can be examined with PBMs to identify subtle differences in the binding profiles of homologous or structurally similar proteins¹⁷. One can search for subtle differences among the moderate and low-affinity *k*-mer-binding sites for related TFs that otherwise share the same high-affinity sites¹⁷. Additionally, by examining the binding specificities of a large number of family members, one can begin to assemble a set of recognition rules for a particular TF structural family, in which the preferred binding sites of individual TFs can be predicted on the basis of the amino-acid identity at discriminatory residues within the protein^{17,41}. Synthetic constructs can also be designed with the aim of engineering novel binding specificity onto an existing scaffold and developing artificial TFs^{42,43}.

Limitations of PBMs

Protein-binding microarrays are limited by the amount of sequence that can be represented on a microarray. Space and technological limitations of early PBMs required the use of separate sets of probe sequences tailored to individual TFs or structural families with previously known sequence preferences^{7,24,25}. Universal PBMs have largely circumvented this problem by using a maximally compact and cost-efficient design⁹; however, for TFs with very long motifs due to an extensive network of protein–DNA contacts, it may be difficult to capture the full range of specificity. This is most problematic for prokaryotic TFs, which tend to dimerize and

may bind to DNA sequences 20 bp or longer. We have made an effort to regularly sample long *k*-mers and gapped *k*-mers in our microarray design, which can help to reconstruct long motifs⁹. Furthermore, the development of higher-density microarrays will enable the coverage of an even greater portion of sequence space. Even so, the construction of a microarray that captures all 12-mers, e.g., requires 16-fold more sequence than an array that captures all 10-mers.

Additionally, as discussed above, the *in vitro* nature of universal PBMs somewhat complicates their use in predicting functional TF-binding sites *in vivo*. Although we have observed good agreement between PBM-derived binding specificities and *in vivo* binding sites, it is impossible to fully replicate the *in vivo* nuclear environment on a microarray. Our standardized protocol uses physiological salt conditions (PBS, pH 7.4) as well as a rank-based statistical analysis framework that is quite robust to the TF concentration used in PBMs; however, different TFs may require different biochemical conditions for optimal binding. In addition, certain TFs may require particular post-translational modifications or protein interaction partners for increased affinity and specificity in DNA binding. The success of a PBM experiment also requires proper expression and folding of the TF under consideration, which is of particular concern when the TF is expressed in a heterologous or *in vitro* system. Consequently, it is difficult to interpret a negative PBM result yielding limited fluorescence intensity. It is also possible that the sequence preferences of an individual TF can be significantly altered by physical interactions with protein cofactors^{44,45} (F. De Masi and M.L.B., unpublished data).

Experimental design

Combinatorial design of universal PBMs. The design of a microarray containing all possible 10-bp sequences in a maximally compact manner has been described previously^{9,11} and is beyond the scope of this paper. Briefly, we have used a de Bruijn sequence of order 10, in which every 10-mer sequence variant is represented exactly once in an overlapping manner. The de Bruijn sequence is partitioned into shorter sequences 36 nt long that are joined to a common 24-nt primer sequence to become the 60-nt probes on the microarray. Each 36-mer contains 27 overlapping 10-mers. Our particular design ensures that all possible contiguous 8-mers and gapped 8-mers up to 12 total positions occur on at least 16 different probes (32 probes when reverse complements are considered) as shown in

Figure 2 | Sequence coverage and redundancy in the ‘all 10-mer’ universal PBM design. **(a)** Each microarray contains four identical subgrids consisting of ~44,000 probes. Every possible 8-mer occurs on at least 16 probes distributed across the subgrid, each time embedded in a different flanking sequence. (For every nonpalindromic 8-mer, its reverse complement occurs on a separate set of 16 probes.) Probes containing the 8-mer CATGGAAA are shown as an example. The common primer sequence at the 3'-end is not shown. **(b)** All possible gapped 8-mers spanning up to 12 total positions are also covered at least 16 times, as shown for the gapped 8-mer CAnTnGnGAAnA.

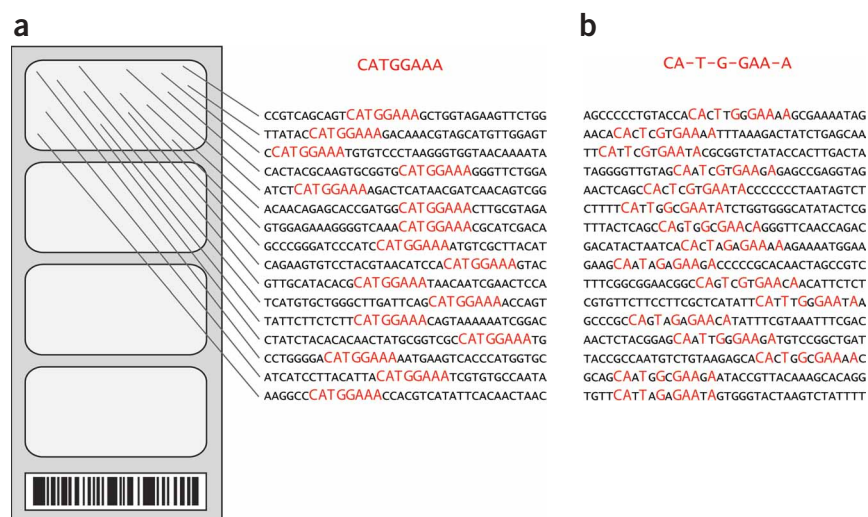


Figure 2. Thus, we are able to reliably estimate the relative preference of a TF for 22.3 million gapped and contiguous 8-mers (4^8 sequence variants of 341 patterns up to 8 of 12) on the basis of a large ensemble of probe intensity measurements. The comprehensive coverage of gapped *k*-mers facilitates the recovery of motifs spanning more than 10 informative positions. Other microarray design strategies are possible; for instance, one may prefer to use an array with tiled genomic sequence endogenous to a particular species. The experimental protocols presented here are suitable for PBM experiments performed on any custom-designed Agilent microarray, as long as the appropriate primer sequence for double-stranding is included. We favor our strategy that uses de Bruijn sequences because it guarantees uniform and compact coverage of all sequence variants, enabling the examination of any TF from any species in an unbiased manner. The flexibility of a design based on de Bruijn sequences is also favorable, as higher-order de Bruijn sequences can easily be adapted for the future construction of higher-density PBMs covering an even greater portion of sequence space, as microarray fabrication technology improves and feature density increases.

Microarray platform options. The protocol described here specifically refers to PBM experiments performed on arrays synthesized by Agilent Technologies. However, we know of no reason why these experiments would not be successful on other microarray platforms, and we expect such deviations would require only relatively minor modifications to the protocol. We have previously created our own smaller-scale, homemade universal PBMs by spotting 8,192 double-stranded oligonucleotide probes that together cover all possible 9-mers (M.L.B. Lab and T.R. Hughes Lab, unpublished data). Other microarray manufacturers, such as NimbleGen, can accommodate custom designs as well. Although the surface chemistries of various microarray slides differ, we have used the PBM protocol described here on multiple slide types without difficulty.

Agilent offers several formats that enable different degrees of multiplexing. Currently, we typically use the ‘4 × 44 K’ format, in which four identical subgrids of ~44,000 probes each can be physically separated into four chambers by a specially manufactured coverslip so that four proteins can be simultaneously

examined on a single slide. Each chamber contains the entire complement of all possible 10-mers. Other currently available formats contain eight chambers (8×15 K') or one chamber (1×244 K') per slide, enabling complete coverage of all 9-mers and all 11-mers, respectively, in each chamber. These numbers are expected to improve as the allowable probe density increases. It should be noted that NimbleGen microarrays could currently accommodate all 12-mers on a single slide. The choice of microarray format depends partly on the number of proteins to be assayed, expectations of the proteins' DNA binding site lengths and cost considerations. For instance, eight-chambered universal PBMs containing all 9-mers potentially offer a more economical choice when multiple proteins are to be examined that are expected to have relatively short motifs.

Protein production options and requirements. DNA-binding proteins can be cloned and expressed by several strategies. We often clone just the DNA-binding domain of a TF, embedded in a modest amount of flanking sequence (often ~ 15 amino acids N- and C-terminal to the DNA-binding domain). Working with smaller polypeptides increases the ease of cloning and protein production as a practical matter; additionally, full-length proteins may possess additional domains that inhibit DNA binding in the absence of interacting protein cofactors⁴⁶. For the TFs for which we have performed a direct comparison, DNA-binding domains and full-length proteins have yielded indistinguishable results on PBMs, or the full-length protein has failed to bind, whereas the domain alone exhibited sequence-specific binding. In contrast, for TFs expected to dimerize (such as helix–loop–helix and leucine zipper proteins), it is necessary to include also known or predicted dimerization domains. Full-length proteins may also be preferable in cases where regions outside of the DNA-binding domain of the TF are expected to confer additional sequence specificity, or if one attempts to assemble heterodimers or protein complexes *in vitro* on PBMs (F. De Masi, M.L.B., unpublished data). For ease of maintenance, sequence verification and transfer into expression vectors for alternate tagging strategies, we typically create a master (donor, or Entry) clone compatible with the GATEWAY⁴⁷ or MAGIC^{17,48} system. We then express each polypeptide as a fusion with glutathione *S*-transferase (GST) at the N terminus. The GST tag can be used for both protein purification and fluorescent labeling of PBMs. Other epitope tags can be used instead, as long as they are compatible with labeling strategies (see below).

Much of our experience is based on expressing fusion proteins in inducible *Escherichia coli* overexpression cultures, followed by purification using glutathione columns or glutathione-coated beads. This has worked quite well for us; however, other expression systems such as mammalian cell culture could be used, especially if there is an indication that particular post-translational modifications may be required. We have also observed that purification from cellular lysate is not always necessary, as only protein that is tagged with GST will produce signal on a PBM that has been stained with fluorophore-conjugated anti-GST antibody⁸. Furthermore, proteins can be expressed by coupled *in vitro* transcription and translation (IVT) reactions using *E. coli* lysate. Clones expressed in *E. coli* and by IVT yield proteins exhibiting identical binding specificities on PBMs in our hands¹⁷. IVT has the potential to dramatically increase the throughput of protein production for large-scale projects, as these reactions can be conducted in parallel

in 96-well plates, take less time than growing overexpression cultures and do not require subsequent protein purification before use of the proteins in PBMs. The PBM protocol described here presumes that the desired epitope-tagged protein has already been produced and that its concentration has been accurately estimated by western blot or another method. PBM experiments are advantageous compared with traditional methods, such as electrophoretic mobility shift assay, in that they require very small quantities of protein, typically just a few hundred nanograms per experiment. Proteins may be stored in a standard buffer (we typically use PBS, pH 7.4, or Tris-HCl, pH 7.0) or as unpurified cellular lysate. We recommend preparing separate aliquots of protein stocks and adding glycerol (final concentration 30% vol/vol) for long-term storage at -80°C . For proteins containing zinc-finger domains, zinc acetate should be added to all protein expression, purification and storage buffers, as indicated in the protocol.

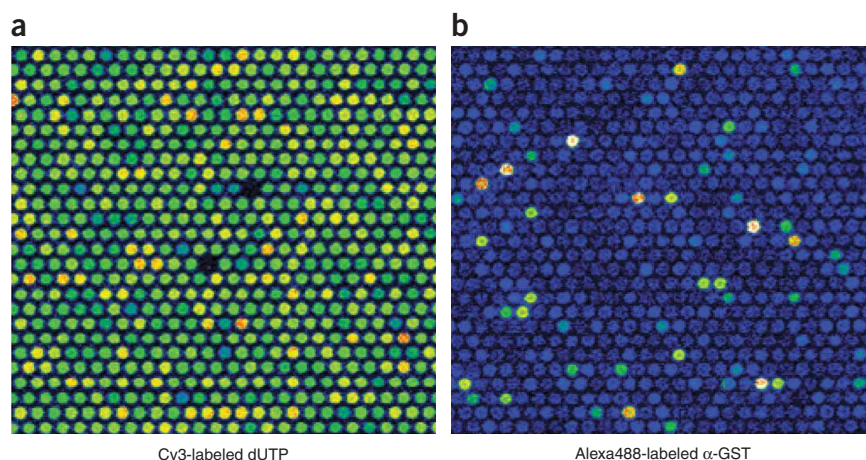
Optimizing primer extension reactions. To use Agilent single-stranded oligonucleotide arrays in PBM experiments, they must first be double-stranded by a solid-phase primer extension reaction. The protocol presented here has been optimized with respect to several parameters, including primer sequence and melting temperature, type of DNA polymerase, fluorescent label conjugated to the nucleotides, concentration of reagents, duration and temperature. This process involved many experiments in which the incorporation of spiked-in fluorescently labeled nucleotides was monitored for a set of specially designed control probe sequences. However, it is possible that the primer extension procedure may be improved further. For example, it is possible that a shorter primer may be used, which would free up additional probe sequence for the inclusion of additional putative binding sites.

These primer extension reactions are quite sensitive to temperature and must be set up rapidly to minimize misannealing of primer and improper double-stranding. Consequently, it is important to monitor the accuracy of each primer extension reaction before using a microarray in a protein-binding experiment. This is accomplished by the addition of small quantities of Cy3-conjugated dUTP to the reaction. The Cy3 signal indicates the amount of double-stranded DNA present at each spot and is used as a normalization factor in the final analysis of the PBM (Fig. 3). This signal reflects the number of adenines in the template strand as well as the sequence context of each adenine; of note, the effect of sequence context varies for different fluorescent tags and polymerases. Therefore, after scanning a primer-extended microarray, we fit the observed signal intensities by a linear regression with 64 parameters, corresponding to every possible trinucleotide preceding each adenine in the template sequence, to ensure that the DNA is properly double-stranded. (The observed and expected Cy3 intensities should exhibit a correlation of $R^2 > 0.7$, as shown in Fig. 4.) We have observed that runs of five or more consecutive guanines are deleterious for primer extension reactions. As a result, we have replaced each probe sequence containing such runs of guanines with its reverse complement.

Selecting optimal protein-binding conditions. We have attempted to devise a single protocol that is best suited to the largest number of TFs in a first-pass experiment. Our protocol uses relatively standard binding conditions (e.g., pH 7.4, $1 \times$ PBS buffer, 100 nM protein). After performing numerous PBM experiments,

Figure 3 | Zoom-in of a universal PBM scan.

(a) Region of a single subgrid, consisting of just over 1% of the total slide area, scanned to detect relative DNA amounts, as indicated by Cy3-labeled dUTP. (b) The same region of the same microarray, scanned with a different laser to detect protein binding, as indicated by Alexa 488-labeled anti-GST antibody. Intensities are shown in false color, with white indicating saturated signal intensity, yellow indicating high signal intensity, green indicating moderate signal intensity and blue indicating low signal intensity.



we believe these conditions to be suitable for most TFs. Furthermore, we specifically use rank-based statistics to analyze PBM data, under the assumption that the ranking of probes by intensity should be invariant to changes in pH or protein concentration even though their relative differences in signal intensity may vary. Nevertheless, the DNA binding of some TFs may be particularly sensitive to salt concentrations or cofactors, and so these buffer conditions should be used in cases when such previous information on preferable alternate conditions is available. For example, zinc should be included in all reactions and wash buffers involving zinc-finger TFs. If a PBM experiment produces faint or background-level signal, it may help to increase the protein concentration, decrease the wash time and stringency and/or alter the binding conditions.

Labeling strategies and scanning considerations. The protocol described here requires that TFs possess a GST tag so that they can be labeled by an Alexa488-conjugated anti-GST antibody (Sigma). Other tagging and labeling methods can theoretically be used. We have successfully used the maltose-binding protein tag and the Flag tag with corresponding fluorescently labeled antibodies in pilot experiments. However, the availability of a commercial fluorophore-conjugated polyclonal anti-GST antibody that results in very bright signal intensity makes GST our tag of choice. **Figure 3** shows a close-up portion of a single microarray, scanned with two lasers to detect DNA concentration, represented by Cy3-labeled dUTP, and protein abundance, represented by Alexa 488-labeled anti-GST antibody. Usage of multiple tags and fluorophores may enable a dual-labeling strategy for comparing the binding specificities of homodimers and heterodimers (or for multiplexing independent TFs) on one microarray, as long as their spectra do not overlap with the fluorescent nucleotides or with each other. Alternatively, TFs could potentially be tagged directly with green fluorescent protein or another fluorescent molecule to eliminate the labeling reaction entirely.

The spot diameter for microarrays manufactured by Agilent is currently $\sim 50 \mu\text{m}$, thus requiring a microarray scanner that is capable of $5\text{-}\mu\text{m}$ resolution scans for accurate image quantification. Higher-density microarrays with smaller feature sizes are anticipated, necessitating even higher resolution scans. Detection of Alexa 488 (488 nm excitation/522 nm emission) requires an argon laser, separate from the Cy3 (543 nm excitation/570 nm emission) and Cy5 (633 nm excitation/670 nm emission) lasers that are part of most standard microarray scanners (including Agilent's own scanner). For our scans, we use a ScanArray 5000 (GSI Lumonics) scanner with an external 488-nm argon laser.

Performing replicate experiments. We frequently perform PBM experiments in duplicate for each TF. Rather than repeat an experiment on a microarray of the same design, however, we use a second microarray with an independent design constructed using a separate de Bruijn sequence of order 10. Our second microarray also contains all possible (nonpalindromic) 8-mers spanning up to 12 total positions on 32 probes each. By combining data from separate microarrays of different designs, we effectively double the number of independent measurements made for every 8-mer, thereby increasing the accuracy. Nevertheless, replicate experiments may not always be necessary. There is substantial redundancy built into our combinatorial microarray design, minimizing the importance of any single probe measurement. For TFs expected to possess short motifs (i.e., 7 or fewer informative nucleotide positions), the sequence coverage provided by a single 'all 10-mer' microarray should be sufficient to capture its full binding specificity. If the aim of an experiment is to compare the binding profiles of two very similar TFs, this can also be accomplished by performing single experiments on the same microarray design¹⁷.

Binding-site representation and analysis strategies. The greatest advantage of universal PBMs, compared with other existing

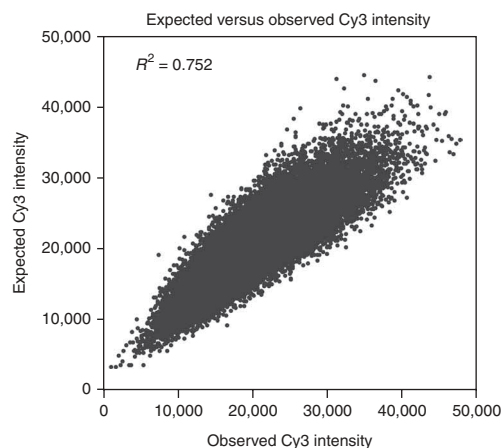
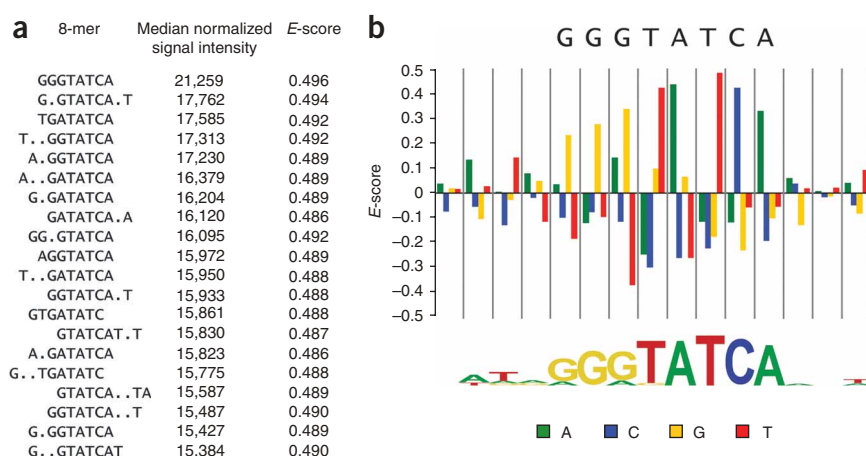


Figure 4 | Correlation between observed and expected Cy3 probe intensities. Expected intensities were determined from sequence, on the basis of the calculated regression coefficients for all trinucleotides.

Figure 5 | Word-by-word and PWM representations of binding specificity. **(a)** Scores for individual k -mers. The top-scoring 8-mers for a PBM experiment using the mouse TF Six6 (see ref. 17) are shown with their corresponding median signal intensities and enrichment scores. The 'median normalized signal intensity' represents the set of ~ 32 probes containing a match to each 8-mer. 'E-score' refers to the enrichment score described in the text. **(b)** Overview of our Seed-and-Wobble method⁹ for motif construction. The top-scoring 8-mer is used as a seed, and the relative preference of each nucleotide variant is systematically tested at each position within and outside the seed. These nucleotide E -scores are converted to probabilities using a Boltzmann distribution and displayed as a sequence logo⁵⁴.



methods for characterizing TF-binding specificities, is that binding to all 'words' up to a given length k is simultaneously assayed. Consequently, these experiments provide a comprehensive look-up table conveying a precise measure of the preference of a particular TF for every sequence variant (Fig. 5a). There are several methods for scoring individual k -mers on the basis of the distribution of signal intensities observed on the microarray. For instance, k -mers can be scored according to the median signal intensity of the set of probes containing each k -mer, which can be further transformed into a Z -score. These measures are useful because they convey information regarding relative differences in DNA occupancy and affinity. However, we have developed a separate rank-based, nonparametric enrichment score (E -score)⁹ that we believe is preferable for a larger number of applications. As the E -score is rank based, it is robust to differences in protein concentration and other binding conditions in the PBM assay. By putting all experiments on the same scale, it enables TFs to be directly compared and data from replicate PBM experiments on different array designs to be easily combined. Finally, the E -score is robust to differences in sample size (i.e., the number of spots harboring a match to a given k -mer), thus providing a uniform standard for comparing palindromes and nonpalindromes and also k -mers of different lengths.

Such comprehensive 'word-by-word' measurements are valuable because they carry information about nucleotide interdependence as well as both high- and low-affinity classes of binding sites, information that is not easily captured in a conventional PWM representation. An exhaustive look-up table can also be used to perform genome-wide scans for potential TF-binding sites. Yet such a list is cumbersome and provides little intuitive feel for the complete binding specificity of the TF. For this reason, and the fact that most existing software for genome scanning for TF-binding sites use PWMs as input⁴⁹, we developed the Seed-and-Wobble algorithm⁹ for PWM construction (Fig. 5b). This approach specifically takes advantage of the unbiased coverage of all k -mers on the array to identify the relative contribution of each base at each position to the binding specificity, and it has proven to be effective at recapitulating the known binding preferences of well-characterized TFs^{9,17}. By making use of the gapped k -mers present in our combinatorial design, Seed-and-Wobble also facilitates the recovery of both gapped motifs and long motifs with more than 10 informative positions. Additional algorithms, such as RankMotif++⁵⁰, Prego⁵¹ and MatrixREDUCE⁵², are similar to Seed-and-Wobble in that they use all binding data rather than assigning an arbitrary cutoff, and they can be applied directly to the normalized data from universal PBM experiments as well.

MATERIALS REAGENTS

- HPLC-purified primer (unmodified) for double-stranding of DNA oligonucleotide array 5'-CAGCACGGACAACGGAACACAGAC-3' (Integrated DNA Technologies)
- High-purity solution dNTPs (GE Healthcare, cat. no. 27203502)
- Cy3-conjugated dUTP (GE Healthcare, cat. no. PA53022)
- Thermo sequenase cycle sequencing kit (USB, cat. no. 78500)
- Tween 20 (Sigma, cat. no. P1379)
- Triton X-100 (Sigma, cat. no. T9284)
- Nonfat dried milk, bovine (Sigma, cat. no. M7409)
- Zinc acetate dihydrate ($\text{Zn}(\text{C}_2\text{H}_3\text{O}_2)_2 \cdot 2\text{H}_2\text{O}$; Sigma, cat. no. Z4540)
- DNA, single-stranded from salmon testes (Sigma, cat. no. D7656)
- Bovine serum albumin (BSA; New England Biolabs, cat. no. B9001S)
- Anti-glutathione S -transferase, rabbit IgG fraction, Alexa Fluor 488 conjugate (Invitrogen, cat. no. A11131)
- Protease, from *Streptomyces griseus* (5.8 U mg^{-1} ; Sigma, cat. no. P6911)
- SDS (Sigma, cat. no. L4390)
- EDTA disodium (Sigma, cat. no. E5134)
- Sodium chloride (NaCl; Fisher, cat. no. S271-10)

- Potassium chloride (KCl; MP Biomedicals, cat. no. 191427)
- Sodium phosphate dibasic (Na_2HPO_4 ; Sigma, cat. no. S7907)
- Potassium phosphate monobasic (KH_2PO_4 ; Sigma, cat. no. P0662)
- Tris base ($\text{C}_4\text{H}_{11}\text{NO}_3$; Fisher, cat. no. BP152-500)
- Magnesium chloride (MgCl_2 ; Sigma, cat. no. M8266)

EQUIPMENT

- Custom 4×44 K microarray, AMADID no. 015681 and/or no. 016060 (Agilent, cat. no. G2514F)
- SureHyb chamber (Agilent, cat. no. G2534A)
- SureHyb gasket cover slides, 1 array per slide (Agilent, cat. no. G2534-60003)
- SureHyb gasket cover slides, 4 array per slide (Agilent, cat. no. G2534-60011)
- Vacuum desiccator (Fisher, cat. no. 086425)
- Hybridization oven (Fisher, cat. no. 1324710)
- Water bath
- Staining dishes (2) and cover (Wheaton Scientific, cat. no. 900303)
- Glass-staining dish slide rack (Wheaton Scientific, cat. no. 900304)
- Magnetic stir plate and stir bars
- Microcentrifuge
- Benchtop centrifuge with microplate rotor (Fisher, cat. no. 0537548)

- Micro slide boxes (VWR, cat. no. 48444-004)
- ScanArray 5000 microarray scanner equipped with argon ion laser (488-nm excitation and 522-nm emission filter; Perkin Elmer)
- GenePix Pro 6.0 microarray analysis software (Molecular Devices)
- Coplin staining jars (VWR, cat. no. 47751792)
- Forceps
- Kimwipes
- Orbital platform shaker
- Syringes with BD Luer-Lok Tip (VWR, cat. no. 309603)
- 0.45- μ m syringe filters (VWR, cat. no. 28196114)
- Lifter Slip coverslips for microarray slides (Fisher, cat. no. 22035809)
- Dust Off XL canned air (VWR, cat. no. 21899080)
- Incubated shaker (New Brunswick Scientific, cat. no. M1352-0004)
- Nalgene disposable sterilization filtration units, 0.2- μ m filter (Fisher, cat. no. 097401A)

REAGENT SETUP

GST-tagged protein Protein can be expressed *in vivo* in *E. coli* cultures, by coupled IVT or by using other expression systems as described above under 'Protein production options and requirements.' Samples may be purified using glutathione beads or columns and eluted in Tris-HCl (pH 7.0) or PBS (pH 7.4), or cellular lysates containing overexpressed GST-tagged protein may be used directly. Add glycerol to a final concentration of 30%. If the protein contains zinc-finger domains, add zinc acetate to a final concentration of 50 μ M. Protein stocks should preferably contain at least 500 nM GST-tagged protein; estimate the protein concentration by western blot and concentrate if necessary. Prepare separate aliquots before freezing for long-term storage at -80°C .

10 \times Thermo sequenase reaction buffer Combine 26 ml of 1 M Tris-HCl, pH 9.5, and 60 ml of sterile water. Dissolve 6.18 g of MgCl_2 and bring final volume to 100 ml using sterile water. Filter-sterilize using a 0.2- μ m Nalgene filter. Store at room temperature (20 – 25°C) for up to 1 year.

10 mM dNTPs Combine 25 μ l each of dATP, dCTP, dGTP and dTTP (all stock solutions at 100 mM) and 900 μ l of sterile water. Vortex to mix. The final mixture contains 10 mM total dNTPs (2.5 mM of each dNTP). Store at -20°C .

1 \times PBS Add 28 g of NaCl, 0.7 g of KCl, 5.04 g of Na_2HPO_4 and 0.84 g of KH_2PO_4 to 3 liters of sterile water. Stir for ~ 30 min on a magnetic stir plate. Add sterile water to bring the final volume to 3.5 liters. Adjust the pH to 7.4 and autoclave to sterilize. (Alternately, 1 \times PBS can be prepared by diluting a stock solution of 10 \times PBS in sterile water.) Store at room temperature.

4 \times PBS Mix 3.2 g of NaCl, 0.08 g of KCl, 0.58 g of Na_2HPO_4 and 0.096 g of KH_2PO_4 with 100 ml of sterile water, adjust the pH to 7.4 and filter-sterilize using a 0.2- μ m Nalgene filter. Store at room temperature.

10% (vol/vol) Triton X-100 Combine 15 ml of Triton X-100 and 135 ml of sterile water. Filter-sterilize using a 0.2- μ m Nalgene filter and store at room temperature.

20% (vol/vol) Tween 20 Combine 30 ml of Tween 20 and 120 ml of sterile water. Filter-sterilize using a 0.2- μ m Nalgene filter, and store at room temperature.

2% (wt/vol) milk blocking solution Dissolve 0.1 g of nonfat dried milk in 5 ml of 1 \times PBS. Allow at least 1 h for milk to enter solution, rotating gently (25 r.p.m.) on an orbital shaker. This can be set up overnight to save time. Filter solution using a syringe and 0.45- μ m filter. Filtered milk can be stored for up to 1 week at 4°C as long as no precipitate forms.

4% (wt/vol) milk blocking solution Prepare as mentioned above (for 2% milk blocking solution), except that 0.1 g of nonfat dried milk should be dissolved in 2.5 ml of 1 \times PBS.

500 \times zinc acetate (25 mM) Dissolve 0.55 g of zinc acetate dihydrate ($\text{Zn}(\text{C}_2\text{H}_3\text{O}_2)_2 \cdot 2\text{H}_2\text{O}$) in 100 ml of sterile water. Filter-sterilize using a 0.2- μ m Nalgene filter and split into 1.5-ml aliquots. Store aliquots at -20°C .

100 \times zinc acetate (5 mM) Combine 200 μ l of 500 \times zinc acetate and 800 μ l of sterile water. Store at -20°C .

PBM wash solution no. 1 Mix 210 ml of PBS and 210 μ l of 10% Triton X-100. If proteins with zinc fingers are being examined, add 420 μ l of 500 \times zinc acetate. **▲ CRITICAL** Prepare fresh on the day of the experiment.

PBM wash solution no. 2 Mix 70 ml of PBS and 350 μ l of 20% Tween 20. If proteins with zinc fingers are being examined, add 140 μ l of 500 \times zinc acetate. **▲ CRITICAL** Prepare fresh on the day of the experiment.

PBM wash solution no. 3 Mix 468 ml of PBS and 12 ml of 20% Tween 20. If proteins with zinc fingers are being examined, add 960 μ l of 500 \times zinc acetate. **▲ CRITICAL** Prepare fresh on the day of the experiment.

PBM wash solution no. 4 Mix 560 ml of PBS and 1.4 ml of 20% Tween 20. If proteins with zinc fingers are being examined, add 1,120 μ l of 500 \times zinc acetate. **▲ CRITICAL** Prepare fresh on the day of the experiment.

PBM stripping solution Combine 68.6 ml of sterile water and 1.4 ml of 500 mM EDTA in a beaker and mix on a magnetic stir plate. Add 7.0 g of SDS and dissolve. Finally, add 0.05 g of protease from *Streptomyces griseus* and dissolve. Continue stirring for 10 min. **▲ CRITICAL** Protease should be stored as a solid powder at -20°C . This stripping solution must be prepared fresh immediately before use.

EQUIPMENT SETUP

Hydration chamber Lift out the tip rack of an empty pipette tip box, fill the bottom of the pipette tip box with about half an inch of sterile water and replace the tip rack. Wipe the inside of the lid and the tip rack with a Kimwipe moistened with 70% ethanol.

PROCEDURE

Double-stranding of Agilent microarrays ● TIMING 3 h

1| Preheat the hybridization oven to 85°C and thaw the primer, dNTPs and Cy3-conjugated dUTP on ice.

2| Prepare the primer extension reaction mixture in an Eppendorf tube using the following reagent volumes. Add the polymerase last. Mix by vortexing, before adding the polymerase enzyme. After adding the polymerase, mix by carefully pipetting up and down and gently inverting the tube. Multiple microarrays can be processed at once.

Reagent	Volume (μ l) per microarray	Final concentration in mixture
Sterile water	775.3	
Thermo sequenase reaction buffer (10 \times)	90	1 \times
Primer (100 μ M)	10.5	1.17 μ M
dNTPs (10 mM total)	14.7	163 μ M
Cy3-dUTP (1 mM)	1.47	1.63 μ M
Thermo sequenase polymerase (4 U μ l $^{-1}$)	8	0.036 U μ l $^{-1}$
Total	900	

3| Prewarm the primer extension reaction mixture, steel SureHyb hybridization chamber(s) and SureHyb gasket cover slide(s) (one chamber per slide) in the hybridization oven at 85°C for 20 min. Be sure that any windows or apertures are covered to prevent photobleaching of Cy3-dUTP.

4| Prewarm the microarray(s) in the hybridization oven at 85°C for 3 min, DNA side up. Microarrays are shipped from Agilent in a vacuum-sealed slide box. Once the seal is broken, unused microarrays should be stored in a vacuum desiccator.

5| Assemble the microarray, primer extension reaction mixture, hybridization chamber and gasket cover slide according to the photographs in Agilent's instruction manual. Place the cover slide face up on the base of the chamber, pipette 900 μ l of reaction mixture onto the center of the cover slide, lower the microarray face down onto the cover slide and fasten the hybridization chamber to seal in the liquid. Return the assembled chamber to the hybridization oven as quickly as possible.

▲ CRITICAL STEP This must be done rapidly to ensure no appreciable drop in temperature of the reagents and equipment. Materials may be removed from the oven, but the oven door should remain closed as much as possible so that the temperature does not decrease. As materials are hot, they should be handled carefully on a lab benchtop. When processing multiple microarrays, assemble each one independently; do not begin the second array until the first has been completely assembled and returned to the oven.

? TROUBLESHOOTING

6| After 10 min at 85 °C, reduce the oven temperature to 75 °C. After 10 more minutes (from when the temperature is changed), reduce the oven temperature to 65 °C. After 10 more minutes, reduce the oven temperature to 60 °C. The gradual decrease in temperature is to ensure proper annealing of the primer to the template DNA. Hold the temperature at 60 °C for 90 min to allow the primer extension reaction to proceed.

7| During the primer extension reaction, prepare 1 liter of wash solution (1 liter of 1× PBS and 1 ml of 10% Triton X-100). Heat 1 liter of wash solution to 37 °C in a water bath.

8| When the primer extension reaction has finished, fill two staining dishes with 500 ml of 37 °C wash solution each. Insert a slide rack and a magnetic stir bar into staining dish no. 1. Remove the microarray chamber from the oven, carefully extract the slide with the sealed gasket cover slide and disassemble it with the slide fully immersed in wash solution in staining dish no. 2 according to Agilent's instructions (i.e., pry apart the microarray and cover slide using the plastic forceps supplied with the steel hybridization chamber). Agitate the microarray in the wash solution and rapidly transfer it to the slide rack in staining dish no. 1. Multiple slides may be washed on the same rack, preferably with the DNA sides facing in toward the center.

9| Wash the microarray by placing the entire staining dish on a magnetic stir plate. Stir at medium speed (generating a small whirlpool) for 10 min. The staining dish can be covered in aluminum foil or an empty inverted ice bucket to reduce photo-bleaching of Cy3-dUTP.

10| Rinse staining dish no. 2 with sterile water, and fill it with 500 ml of 1× PBS at room temperature. Rapidly transfer the entire slide rack and magnetic stir bar to staining dish no. 2. Stir at medium speed on a magnetic stir plate for 3 min.

11| Remove the slide rack from the wash solution (slowly remove for ~10 s for uniform drying).

? TROUBLESHOOTING

12| Centrifuge the microarray(s) in a slide box for 1 min at 500 r.p.m. (40g) to dry at room temperature.

13| Scan the microarray(s) with at least 5- μ m resolution detection using a laser and filters suitable for Cy3 (excitation 543 nm, emission 570 nm). Use laser power settings such that all spots are significantly above background signal intensity levels but that no spots exhibit saturated signal intensities. Save the scanned images as TIF files. An example of Cy3 scan is shown in **Figure 3a**.

■ PAUSE POINT Double-stranded microarrays can be stored in a slide box in the dark at ambient conditions for weeks before use in a protein-binding experiment.

? TROUBLESHOOTING

Protein-binding and labeling reactions ● TIMING 5 h

14| Prepare blocking solutions of 2% and 4% milk (wt/vol) dissolved in PBS, and PBM wash solutions nos. 1–4, described in REAGENT SETUP. Thaw all materials needed for the PBM experiment on ice: zinc acetate, BSA, salmon testes DNA and GST-tagged protein.

15| Prewet a double-stranded microarray in 70 ml of PBM wash solution no. 1 in a Coplin jar for 5 min, stirring at 125 r.p.m. on an orbital shaker. Up to three PBM slides can be processed in parallel in a single Coplin jar. We suggest that no more than three PBM slides be processed at any one time, although an experimentalist may further stagger experiments to allow increased PBM slide processing after gaining comfort with the protocol.

16| Remove the microarray from PBM wash solution no. 1. Dry the back and edges with a Kimwipe. Pipette 150 ml of 2% milk blocking solution, drop by drop, over the printed area of the microarray. Slowly place a Lifter Slip coverslip onto the microarray to uniformly distribute the blocking solution, being careful to avoid bubbles.



PROTOCOL

17| Incubate the microarray and 2% milk blocking solution at room temperature for 1 h in a hydration chamber (see EQUIPMENT SETUP). Store the chamber in the dark to avoid excessive photobleaching of the labeled DNA.

18| During the blocking step, prepare protein-binding mixtures for each chamber of the PBM, including BSA and salmon testes DNA as nonspecific protein and DNA competitors, respectively. Four-chambered '4 × 44 K' microarrays (described here) can hold a volume of 175 µl in each chamber. For eight-chambered '8 × 15 K' microarrays, volumes should be proportionately scaled down to a total of 75 µl. Mix carefully by pipetting up and down or flicking the tube with a finger. Store protein-binding mixtures at room temperature for at least 30 min before applying them to a microarray.

Reagent	Volume (µl) per chamber	Final concentration in mixture
Sterile water	Varies (to 175 total)	
Zinc acetate (100×) ^a	1.75	1×
4× PBS	21.9	0.5×
4% milk blocking solution	87.5	2% (wt/vol) milk
BSA (10 mg ml ⁻¹)	3.5	0.2 mg ml ⁻¹
Salmon testes DNA (53 µg ml ⁻¹)	1.0	0.3 µg ml ⁻¹
GST-tagged protein	Varies	100 nM
Total	175	

^aZinc acetate is necessary only for proteins containing zinc-finger domains.

19| Fill staining dish no. 1 with 500 ml of 1× PBS. (Add zinc acetate to a final concentration of 50 µM if zinc-finger proteins are being examined.) This will be needed multiple times throughout the experiment and should be kept covered when not being used.

? TROUBLESHOOTING

20| After blocking for 1 h, gently slide off the Lifter Slip coverslip along the length of the microarray at the short end and wash the microarray in 70 ml of PBM wash solution no. 2 in a Coplin jar for 5 min, stirring at 125 r.p.m. on an orbital shaker.

21| Transfer the microarray to a separate Coplin jar filled with 70 ml of PBM wash solution no. 1 using metal forceps and wash for 2 min at 125 r.p.m. on an orbital shaker.

22| During the washes, prepare the steel SureHyb hybridization chamber and four-chambered SureHyb gasket cover slide according to Agilent's instructions. Place the cover slide face up on the base of the chamber and pipette 175 µl of protein-binding mixture onto the center of each chamber, as shown in **Figure 6**. Note carefully which protein was added to which chamber.

23| Remove the microarray from PBM wash solution no. 1 and rinse it briefly in staining dish no. 1 to remove excess detergent from the slide surface. (Submerge the microarray in the staining dish and remove it slowly over the course of ~10 s, tilted slightly face down.) The microarray should be dry upon removal.

24| Lower the microarray face down onto the gasket cover slide, being careful to prevent leakage of protein binding reaction mixture from one chamber to another (**Fig. 6**). Immediately assemble and tighten the steel hybridization chamber. If bubbles form, they can be moved outside of the DNA subgrid by gently tapping the steel chamber against a hard surface.

? TROUBLESHOOTING

25| Incubate the chamber with protein-binding mixture at room temperature for 1 h in the dark, sitting flat.

26| During the protein-binding step, prepare the fluorophore-conjugated antibody mixture (1:40 dilution of Alexa 488-conjugated anti-GST (Invitrogen) in 2% milk blocking solution). Prepare a total of 800 µl for each microarray. Mix carefully by pipetting up and down or briefly vortexing. Store the antibody mixture at room temperature in the dark for at least 30 min, until Step 30.

Reagent	Volume (µl) per microarray
2% milk blocking solution	778.4 (or 780) ^a
Alexa 488-conjugated anti-GST (Invitrogen, cat. no. A11131)	20
Zinc acetate (500×) ^a	1.6 (or 0) ^a
Total	800

^aZinc acetate is necessary only for proteins containing zinc-finger domains.

27| Fill staining dish no. 2 with 400 ml of PBM wash solution no. 3.

28 | When the protein-binding reaction is finished, carefully extract the microarray and sealed gasket cover slide from the hybridization chamber, and disassemble it while immersed in PBM wash solution no. 3 in staining dish no. 2 as mentioned above (pry apart the microarray and cover slide using plastic forceps). Agitate the microarray in the wash solution, and rapidly transfer it to a Coplin jar already filled with 70 ml of wash solution no. 3. Wash for 3 min at 125 r.p.m. on an orbital shaker.

▲ CRITICAL STEP If not done properly, this step can lead to uneven signal on the microarray. Pry apart the microarray and cover slide quickly. Let the cover slide fall to the bottom of the staining dish. Shake the microarray underwater vigorously, allowing the contents of the chamber to disperse. Transfer the microarray from the staining dish to the Coplin jar quickly to minimize drying, as it is exposed to air.

29 | Transfer the microarray to a separate Coplin jar filled with 70 ml of PBM wash solution no. 1 using metal forceps, and wash for 2 min at 125 r.p.m. on an orbital shaker.

30 | During the washes, rinse the gasket cover slide with distilled water while gently rubbing the surface with gloved fingers to remove any particles. Rinse again with 70% ethanol, and dry the cover slide with Dust Off canned air. Place the cover slide face up on the base of the steel hybridization chamber and dispense 175 μ l of antibody mixture onto the center of each chamber.

31 | Remove the microarray from PBM wash solution no. 1 and rinse it briefly in staining dish no. 1 to remove excess detergent from the slide surface and dry the slide, as mentioned in Step 23.

32 | Lower the microarray face down onto the gasket cover slide (**Fig. 6**). Immediately assemble and tighten the steel hybridization chamber. If bubbles form, they can be moved outside of the DNA subgrid by gently tapping the steel chamber against a hard surface. Incubate the chamber with antibody mixture at room temperature for 1 h in the dark to prevent photobleaching of the Alexa 488 antibody.

? TROUBLESHOOTING

33 | Rinse staining dish no. 2 and fill it with 400 ml of PBM wash solution no. 4.

34 | When the antibody labeling reaction is finished, carefully extract the microarray and sealed gasket cover slide from the hybridization chamber and disassemble it while immersed in PBM wash solution no. 4 in staining dish no. 2 as mentioned above. Agitate the microarray in the wash solution and rapidly transfer it to a Coplin jar already filled with 70 ml of wash solution no. 4. Wash for 3 min at 125 r.p.m. on an orbital shaker.

▲ CRITICAL STEP As mentioned in Step 28, this must be done quickly to ensure signal uniformity and minimize drying as the microarray is exposed to air.

35 | Transfer the microarray to a separate Coplin jar filled with 70 ml of PBM wash solution no. 4 and wash for 3 min at 125 r.p.m. on an orbital shaker.

36 | Transfer the microarray to a separate Coplin jar filled with 70 ml of 1 \times PBS and wash for 2 min at 125 r.p.m. on an orbital shaker. (Add zinc acetate to a final concentration of 50 μ M if zinc-finger proteins are being examined.)

37 | Rinse the microarray briefly in staining dish no. 1 to remove excess detergent from the slide surface and dry the slide, as mentioned in Step 23. Centrifuge the microarray in a slide box for 1 min at 500 r.p.m. (40g) at room temperature to remove all liquid.

■ PAUSE POINT Microarrays can be stored at room temperature in a slide box in the dark at ambient conditions for weeks at a time before scanning without any appreciable loss in Alexa 488 signal. Other fluorophores may be less stable, however.

? TROUBLESHOOTING

38 | Scan the microarray with at least 5- μ m resolution detection using a laser and filters suitable for Alexa 488 (excitation 488 nm, emission 522 nm). Take a series of scans at multiple laser power settings (keeping the photomultiplier tube (PMT) gain

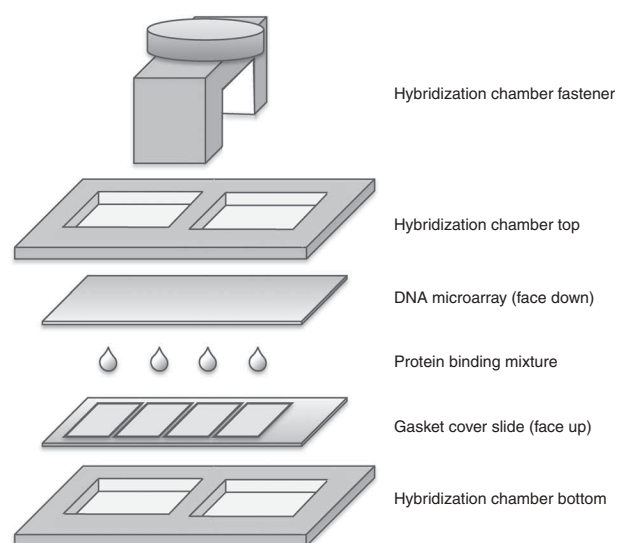


Figure 6 | Schematic of Agilent SureHyb hybridization chamber for protein-binding reactions. The gasket cover slide, protein binding mixture and microarray are sandwiched between both halves of the steel hybridization chamber. A four-chambered cover slide is used for the protein-binding and antibody-labeling incubations, whereas a single-chambered cover slide is used for primer extension. This figure is not drawn to scale.

fixed) to ensure reliable measurements over a large dynamic range of intensities. Ensure that there is at least one scan for which no spots exhibit saturated signal intensities (median pixel intensity = 65,536). The lowest power scan should display the brightest spots at subsaturated signal intensities, and the highest power scan should display the faintest spots at above-background signal intensities. Owing to the fact that four arrays are printed on each slide, a series of 4–5 scans may be necessary to capture the full dynamic range in all four chambers. Save the scanned images as TIF files. An example of Alexa 488 scan is shown in **Figure 3b**. (We note that alternate microarray scanners, such as the Axon GenePix scanner, may be used at this step and that different intensity scans may be obtained by holding the laser power fixed and varying the PMT gain. The user should consult the instruction manual for the particular scanner being used.)

? TROUBLESHOOTING

Protease digestion for subsequent reuse of PBMs

39| Prepare 70 ml of PBM stripping solution, mixing for at least 10 min on a magnetic stir plate. This is enough to fill one Coplin jar, which can hold up to three microarrays.

40| Place up to three protein-bound microarrays into a Coplin jar filled with 70 ml of stripping solution. Wash overnight (~16 h) at 37 °C in an incubated shaker at 200 r.p.m., fastened or taped to the base of the shaker to prevent tipping.

41| Transfer the microarray(s) to a Coplin jar filled with 70 ml of PBM wash solution no. 3, and wash for five min at room temperature on an orbital shaker at 125 r.p.m.

42| Repeat for two more washes in PBM wash solution no. 3 for 5 min each.

43| Wash the microarray(s) in a Coplin jar filled with 70 ml of 1× PBS for 2 min at room temperature on an orbital shaker at 125 r.p.m.

44| Rinse the microarray(s) briefly in a staining dish filled with 500 ml of PBS to remove excess detergent from the slide surface and dry the slide(s), as mentioned in Step 23. Centrifuge the microarray(s) in a slide box for 1 min at 500 r.p.m. (40g) at room temperature to remove any residual liquid.

45| Scan the microarray at the highest laser power settings using lasers suitable for Cy3 and Alexa 488 to ensure that there is no appreciable loss in DNA signal (Cy3), but that all protein signal has been removed (Alexa 488).

Image analysis and data normalization

46| Using GenePix Pro version 6.0 software, compute the background-subtracted probe signal intensities for all scanned image files. This requires a GenePix Array List file containing information regarding the coordinates and median pixel identities of all spots within each subgrid, which is supplied by Agilent with each microarray. Align each block of spots over the corresponding subgrid, and manually flag problematic spots as ‘bad’ (i.e., spots with obvious scratches, dust flecks and so on). Save the intensities for each subgrid as a separate GenePix Results (GPR) file.

▲ CRITICAL STEP There are several software packages and strategies available for analyzing and normalizing microarray data. The remainder of this protocol describes our recommended approach, although variants of these methods may be acceptable. We have written Perl scripts to conduct many of the analyses described below. These programs, demo files and a complete documentation and explanation are available in the ‘Software’ section of the Bulyk Lab website: http://the_brain.bwh.harvard.edu/software.html. In particular, Steps 48 through 53 can be executed using the program ‘normalize_agilent_array.pl’.

47| Combine the results from the series of Alexa 488 scans taken at multiple laser power settings (and constant PMT gain) using masliner software⁵³. Masliner performs a linear regression, using the low-power scans to resolve the relative intensity differences among the brightest spots exceeding saturation levels in the high-power scans (**Fig. 7**). Masliner can be downloaded for free to academic users from the following URL at the Church Lab website: <http://arep.med.harvard.edu/masliner/supplement.htm>.

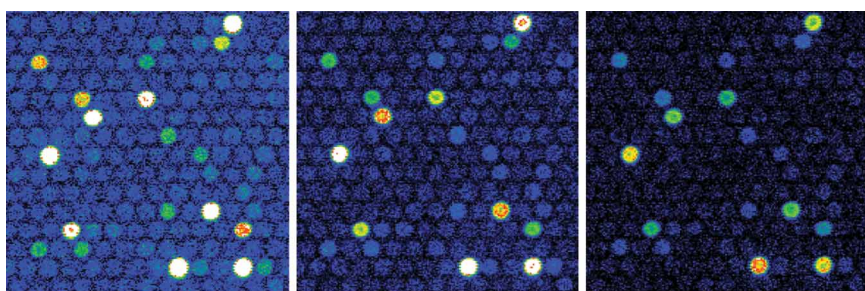


Figure 7 | Replicate scans at multiple laser power settings for integration by masliner⁵³. The same portion of the same microarray is displayed for three scans at varying laser power settings. (The color scheme is the same as for **Fig. 3**.) The dimmest scan (right) can be used to resolve relative differences in signal intensity for spots with saturated intensities in the brightest scan (left), whereas the brightest scan provides above-background signal intensities for spots with low signal intensity.

48| Discard individual features flagged as ‘bad’ in the DNA scan (Cy3) or in any of the PBM scans (Alexa 488). Separately, remove all Agilent and user-defined control spots, leaving only those probes derived from the original de Bruijn sequence (41,944 in total).

? TROUBLESHOOTING

49| Calculate regression coefficients representing the contribution of each trinucleotide in the probe sequence to the total Cy3 signal intensity by performing a linear regression using the remaining probe sequences and signals. (For this calculation, use only the part of the probe sequence that is downstream of the primer.) This step is necessary because we have found the incorporation of Cy3-dUTP to be dependent on the local sequence context of each adenine in the template strand. The regression can be performed using our software described above.

50| Use these coefficients to compute the expected Cy3 signal of each probe on the basis of its DNA sequence (**Fig. 4**). Discard probes with observed Cy3 signals twofold smaller or larger than expected.

? TROUBLESHOOTING

51| Normalize the Alexa488 signal of each probe by dividing by the ratio of its observed-to-expected Cy3 signal.

52| Compute the median Cy3-normalized Alexa 488 intensity over all spots in the entire grid (global median) and also for the ‘local neighborhood’ of each spot (local median; 15×15 block centered on each spot). Divide the normalized Alexa 488 signal at each spot by the ratio of the local median to the global median. (For spots near the margins of the grid, use the 15×15 block of spots along the edge to represent the local neighborhood.)

53| Rank all probe sequences in descending order according to their Cy3-normalized, spatially adjusted Alexa 488 signal intensities.

Sequence analysis

54| Every possible 8-mer occurs on at least 32 different probes (except for palindromes, which occur on 16 probes). This applies to contiguous 8-mers as well as all gapped 8-mers spanning up to 12 positions (e.g., CA-T-G-GAA-A), as illustrated in **Figure 2**. For each contiguous and gapped 8-mer, compute the median signal intensity over the set of ~ 32 or ~ 16 probes in which it occurs. (For this analysis, consider only the part of the probe sequence that is downstream of the primer.) We have observed these median signal intensity values to be roughly proportional to the relative affinities for these sequences⁹.

▲ CRITICAL STEP The main aim of this analysis is to transform the signal intensities of probes (each of which is composed of several overlapping 8-mers) into scores reflecting the relative preferences for 8-mers (each of which occurs on several different probes). As above, we have written Perl scripts to conduct the analyses described in Steps 54 through 63. These programs and a thorough explanation of their proper usage are available in the ‘Software’ section of the Bulyk Lab website: http://the_brain.bwh.harvard.edu/software.html. In particular, Steps 54 through 61 can be executed using the program ‘seed_and_wobble.pl’.

55| Separately, using the probe rankings, compute the enrichment score (E -score)⁹ for each 8-mer. Define the ‘foreground’ features as those containing a match to the 8-mer and define the ‘background’ features as all others. Considering the brightest 50% of features in both the foreground and the background, the E -score corresponds to the geometric area between the foreground and background detection rate curves. Mathematically, this is expressed as $(\rho_B/B - \rho_F/F)/(B + F)$, where B and F are the sample sizes of the background and foreground, respectively, and ρ_B and ρ_F are the sums of their respective ranks. The E -score ranges from -0.5 (lowest enrichment) to $+0.5$ (highest enrichment) and is approximately equal to the area under the receiver operating characteristic curve (AUC) minus 0.5 (see ref. 9).

56| Choose the 8-mer (contiguous or gapped) with the largest E -score as a seed for constructing a compact PWM representation of the protein’s binding specificity.

? TROUBLESHOOTING

57| At each position within the seed, compute the ‘reduced E -score’ for each of the four nucleotide variants. For the reduced E -score, the foreground consists of the ~ 32 probes containing the 8-mer with the nucleotide under consideration at that position, and the reduced background consists of the ~ 96 probes containing an 8-mer with any of the other three nucleotides. All probes belonging to the foreground and reduced background are considered. Mathematically, this calculation is the same as above.

58| Transform the reduced E -scores to probabilities using a Boltzmann distribution. This can be achieved with the formula: $P(j) = e^{\gamma * E_j} / \sum e^{\gamma * E_{j'}}$ for $j' = \{A, C, G, T\}$ and $\gamma = \ln(10,000)$. E_j represents the E -score for base j . We use $\ln(10,000)$ as a scaling factor to calibrate the probability distribution such that an E -score of 0.5 corresponds to a probability of 0.99 (see ref. 9).

59 Identify the least informative position within the 8-mer seed as the position with the minimum relative entropy $H(p) = \sum P(j, p) \times \log_2(P(j, p)/0.25)$.

60 Discard the least informative position within the seed (above), and consider every additional position outside the seed that, when combined with the remaining seven positions, constitutes a pattern whose 4^8 sequence variants are also covered on at least 32 probes each. For the 'all 10-mer' microarray designs described here, this includes all gapped 8-mers spanning up to 12 positions, as well as many longer patterns. A complete list of 8-mer patterns covered by our design can be found at the Bulyk Lab website: http://the_brain.bwh.harvard.edu/software.html.

61 At each of these new positions outside the original seed, compute the reduced E -scores for all four nucleotide variants and transform these into probabilities as described above. The calculated probabilities can be represented as a matrix of N columns (i.e., nucleotide positions of the binding site) by four rows (corresponding to A, C, G and T).

62 The resulting PWM can be graphically displayed as a sequence logo using a variety of Web-based programs, such as enoLOGOS⁵⁴ (<http://chianti.ucsd.edu/enologos/>). An overview of the PWM construction process is illustrated in **Figure 5b**.

63 If two PBM experiments were performed on different microarray designs, calculate combined E -scores for each 8-mer by averaging their E -scores from each individual microarray. Choose the 8-mer with the largest average E -score as a seed. Compute a matrix of reduced E -scores for each separate microarray using the chosen seed, and average the reduced E -scores to construct a single combined matrix. Transform these into probabilities as described above. (This step can be executed using the program 'seed_and_wobble_two_array.pl' on the Bulyk Lab website.)

● TIMING

Protein-binding microarray experiments are very rapid. Double-stranding and protein-binding reactions can be performed either on the same day or on different days. Two to three PBM slides can be processed in parallel for both stages. When performing a series of PBM experiments, much of the data normalization and sequence analysis for the first set of PBMs can be completed during the long incubation steps during the next set of experiment(s).

Steps 1–13, double-stranding Agilent microarrays: 3 h

Steps 14–38, protein binding and antibody staining of protein-bound arrays: 5 h

Steps 39–45, protease digestion: overnight incubation, followed by 1 h of washes and scanning

Steps 46–53, image analysis and data normalization: 1–3 h

Steps 54–63, sequence analysis: 1–2 h, using the software we provide at the Bulyk Lab website

? TROUBLESHOOTING

Step 5

A total of 900 μ l of primer extension reaction mixture should completely fill the volume of the SureHyb gasket cover slide. We routinely reuse these cover slides 20 or more times. However, if significant leakage of liquid occurs or if a seal does not properly form between the cover slide and microarray, it may be necessary to replace the cover slide.

It is important to execute this step rapidly to avoid a significant drop in temperature. If the reagents are not maintained at close to 85 °C, improper double-stranding may occur due to primer misannealing and/or formation of secondary structures in the template strand. This will be reflected in the quality of the fit (R^2) between the observed and expected Cy3 probe intensities in Step 50.

Step 11

Owing to the hydrophobic surface properties of Agilent slides, the microarray(s) should be mostly dry after removal from 1× PBS. If there are any droplets remaining, these can leave tracks behind during the centrifugation in Step 12. Excess liquid can be removed by dabbing the edges and back of the microarray with a Kimwipe. If the printed area of the microarray is still noticeably wet, rinse the microarray again in 1× PBS and remove it slowly over the course of ~10 s, tilted slightly face down.

Step 13

If the signal is uneven, the washes may need to be performed more vigorously. If there are speckles and dust particles visible in the scan, make sure that all containers and vessels used to store and prepare the wash solutions are cleaned thoroughly. Wash solutions can also be filtered before use.

The overall fluorescence intensity should be very bright if this protocol is followed as written. If for some reason the spots are barely visible at the highest laser power settings, possible improvements include using more Thermo sequenase polymerase, more Cy3-labeled dUTP and/or less unlabeled dNTP. (However, if the ratio of labeled dUTP to unlabeled dTTP exceeds ~5%, the Cy3 conjugate may significantly interfere with TF-DNA binding.) Take precautions to store all fluorescent materials in the dark to

avoid photobleaching. It is also advisable to double-check that the proper laser and filter settings are being used by the microarray scanner.

Step 19

If the staining dish is not kept covered (or if it is not thoroughly rinsed before use), dust or other particles may enter the wash solution. This can lead to speckles interfering with particular probe measurements during the scanning and image analysis.

Step 24

Spillover between adjacent chambers may occur if the microarray is not dry after the wash in $1\times$ PBS in Step 23. (Excess liquid can be removed by dabbing the edges and back of the microarray with a Kimwipe after Step 23.) A 175- μ l protein-binding mixture should just barely fill the volume of the gasket cover slide without leakage; however, the volume of the binding mixture can be reduced even further if spillover becomes a problem. The steel hybridization apparatus should be assembled and tightened quickly for the protein mixture to spread out throughout each chamber in the cover slide and for a seal to form. (If this occurs too slowly, the signal within a chamber may not be perfectly uniform.) It is important to check for bubbles after assembling the hybridization chamber. If bubbles are not moved to the side, the affected probes will have to be flagged and removed from the analysis.

Step 32

As mentioned in Step 24, drying the microarray prevents spillover between adjacent chambers. If the microarray and coverslip are not assembled quickly enough, the center of each subgrid may appear brighter than the margins due to the uneven spread of fluorescently labeled antibody throughout the chamber. As before, if bubbles are not moved to the side, the probes on corresponding area of the slide will exhibit little to no signal intensity.

Step 37

As mentioned in Step 11, the hydrophobic surface properties of Agilent slides should leave the microarray mostly dry after removal from $1\times$ PBS. If there are any droplets remaining, these can leave tracks behind during centrifugation. Excess liquid can be removed by dabbing the edges and back of the microarray with a Kipwipe.

Step 38

A successful PBM experiment will exhibit a broad range of signal intensities, with the brightest probes being visible at moderate laser power settings (50–75% laser power). If all probes are faint at even the highest laser power settings, this most likely reflects a problem with the PBM experiment and may present further problems in the subsequent motif discovery steps. The experiment may have failed due to misfolded protein, improper binding buffer conditions or the absence of required protein cofactors or post-translational modifications. These problems can be addressed only by altering the conditions for protein expression and/or protein binding. However, it is possible that the protein does bind DNA sequence specifically but with low affinity or with a fast dissociation rate. In this case, the signal can be increased by repeating the PBM experiment with a higher protein concentration, a higher antibody concentration and shorter wash times.

If problems continue, we suggest attempting a new PBM experiment with the *S. cerevisiae* TF Cbf1. We have found this protein to be easily expressed in, and purified from, *E. coli* and robust in our protocols for protein binding experiments. The resulting scan should exhibit a broad range in probe signal intensities, with a modest number of extremely bright probes. Sequence-verified full-length *S. cerevisiae* CBF1 cloned into the Gateway Entry vector pDONR201 is available (Cbf1 pDONR201, CloneID ScCD00009385) through the PlasmID repository at <http://plasmid.med.harvard.edu/PLASMID>.

Step 48

Some proteins may exhibit a high degree of nonspecific binding to single-stranded DNA. In such cases, the Agilent control probes, which are not double-stranded by primer extension, may be among the brightest spots on the microarray. Therefore, it is important to always filter out these spots before sequence analysis.

Step 50

The observed and expected Cy3 signal intensities should always exhibit a reasonably high correlation ($R^2 > 0.7$). If instead, $R^2 \approx 0$, check to make sure that the GenePix Array List file contains the correct information for the microarray design that was used and that it was correctly aligned with the grid of spots in GenePix Pro. Probes that are problematic during primer extension will exhibit Cy3 signal intensities much lower than expected. (We had originally observed this for template strands containing long runs of guanine. Consequently, all probe sequences with five or more consecutive guanines have since been replaced in our Agilent array designs by their reverse complements.)

Step 56

Occasionally, the method for PWM construction outlined in Steps 56–61 may fail for TFs with exceptionally long motifs. This is particularly problematic for prokaryotic TFs, which frequently dimerize and bind to DNA sequences as long as 20 bp. This is because the most significant gapped 8-mer may occur in an unfavorable sequence context in the majority of its ~32 occurrences. In such cases, it may be possible to recover a specific PWM using a conventional motif finder by taking the sequences from the top *N* brightest spots as input⁸. This is not an optimal approach as it requires setting an arbitrary threshold above which all sequences are treated equally; however, it can occasionally lead to the successful recovery of the appropriate motif when the method outlined here fails. For example, MultiFinder integrates several previously developed motif discovery algorithms and can be used for this purpose⁵⁵.

ANTICIPATED RESULTS

Expected final results

Figure 3b shows a portion of a scan from a representative PBM experiment. All probes are usually visible above background fluorescence levels (i.e., between spots, where there is no DNA), but there is often a broad range in probe signal intensities. The majority of probes are typically relatively faint with similar signal intensities, corresponding to nonspecific binding of protein. The remaining probes show evidence of specific binding, often with a small fraction of them exhibiting very high intensities. These probes contain the highest affinity binding sites. PBMs exhibiting such a broad distribution of signal intensities nearly always produce high-quality binding data and very high *k*-mer *E*-scores (i.e., $E \geq 0.45$). Furthermore, sometimes the PBM data with seemingly uniform distributions of probe intensities will produce significant *E*-scores and PWMs with high information content as well. As our scoring method is based on rank-order statistics, it is the relative ordering of probes and not the magnitude of their signal intensity differences that determines the degree of enrichment of a particular *k*-mer or motif. Consequently, it is always necessary to conduct a full analysis of each experiment before concluding that there was no sequence-specific binding. Occasionally, a PBM experiment will fail to produce a significant motif, either because the Alexa 488 signal intensity (i.e., that attributable to protein binding) is too faint or because all probes appear to exhibit the same degree of (nonspecific) binding. As described above, it is difficult to interpret a negative result, as it could be due to misfolded protein, improper binding buffer conditions or the absence of required protein cofactors or post-translational modifications. For many of these cases, it may be necessary to repeat the experiment under different conditions to achieve the desired results. Nevertheless, in large-scale screens that we have conducted, we have observed a success rate between 40% and 50% for proteins produced in *E. coli* or by coupled IVT and tested in a single pass at 100 nM in the standard binding conditions described here.

Evaluating data quality and calculating significance

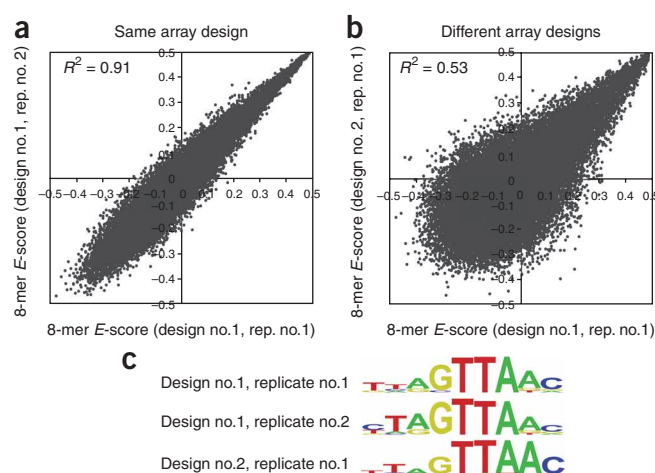
The success of a PBM experiment can be estimated qualitatively by the overall distribution of Alexa 488 signal intensities observed in the scan. However, the quality of the binding data can truly be judged only by examining the *k*-mer *E*-scores derived from the preceding analysis. One indicator of a successful experiment is the occurrence of many *k*-mers with high *E*-scores. Our criterion for concluding that a protein exhibits specific binding is the observation of at least one 8-mer with an *E*-score > 0.45 ; however, most high-quality experiments produce a maximum *E*-score > 0.49 . In a survey of 168 mouse homeodomain TFs, we found, on average per TF, 146 contiguous 8-mers with $E > 0.45$ and 15 with $E > 0.49$ (see ref. 17). A second indicator of a successful experiment is that most of the top-scoring *k*-mers resemble each other and are easily aligned. The motifs of sequence-specific TFs typically tolerate degeneracies at some nucleotide positions of their binding sites. Consequently, the presence of high-scoring 8-mers that contain single mismatches or offsets with respect to each other bolsters the confidence that these 8-mers represent true TF-binding sites, especially considering that each 8-mer score is based on measurements from an independent set of 32 probes.

It is often informative to compute the statistical significance of a particular *E*-score in a PBM experiment. We have calculated the distribution of 8-mer *E*-scores from negative control experiments performed using free GST (rather than GST-tagged TF) and used these to estimate the false discovery rates at various *E*-score thresholds (data not shown). Depending on the TF and the number of 8-mers surpassing each threshold, a false discovery rate of 0.01 typically corresponds to *E*-scores of approximately 0.32–0.36. Calculating significance in this manner enables us to determine the total number of likely true positive binding site sequences for a given TF.

Reproducibility across different array designs

For PBM experiments performed with the same protein on separate ‘all 10-mer’ microarray designs, we observe highly consistent 8-mer *E*-scores. As shown in **Figure 8**, the correlation among 8-mer *E*-scores is also high for experiments performed on different microarray designs⁹. Furthermore, the combined data (from averaging across separate arrays) are often more accurate because they are based on twice as many independent measurements⁹. (For TFs with short motifs (i.e., seven or fewer informative nucleotide positions), the benefits of replicate experiments with multiple microarray designs are reduced because a single

Figure 8 | Correlation in 8-mer enrichment scores obtained from replicate experiments. **(a)** Scatter plot comparing 8-mer scores from two PBM experiments using the mouse TF Tcf1 (see ref. 17) performed on microarrays of the same 'all 10-mer' design. **(b)** Scatter plot comparing 8-mer scores from two PBM experiments using Tcf1 performed on microarrays of complementary 'all 10-mer' designs. *E*-scores for significantly bound 8-mers are consistent among all replicate experiments. **(c)** Sequence logos representing PWMs derived for each data set.



experiment is typically sufficient.) This increase in accuracy can be understood by considering the sources of variability in probe signal intensity. The same *k*-mer may lead to somewhat different signal intensities on different spots owing to its orientation and position on the probe relative to the slide surface⁹. Additionally, two probes with the same *k*-mer may exhibit different signal intensities due to different flanking sequences, both proximal (which may influence binding affinity to the *k*-mer) and distal (which may contain additional binding sites of various affinities). For these reasons, our *k*-mer scoring method relies on multiple measurements from a large ensemble of spots (at least 32 spots for each nonpalindromic 8-mer, and at least 16 spots for each palindromic 8-mer). Nevertheless, in a given array design, a particular *k*-mer may frequently occur close to (or far from) the slide surface or may happen to fall on the same probe as a strong binding site more times than expected by chance. By doubling the number of independent measurements, we further minimize these sources of variation. This has the greatest impact on *k*-mers with *E*-scores near 0. The artificially high correlation across the entire range of *E*-scores in **Figure 8a** can be explained by systematic effects that are fixed within a single array design. **Figure 8b** shows that *E*-scores < 0.2 are in the realm of noise but that higher *E*-scores are very consistent across separate array designs.

Occasionally, the correlation in the *E*-score scatter plot for a pair of PBM experiments may not be as strong as shown in **Figure 8**. For example, one experiment may produce significantly fewer *E*-scores above any given threshold. This is indicative of a noisy data set and can usually be detected in the scanned image itself. In such cases, it is preferable to rely on data from a single array rather than to combine a high-quality data set with a noisy data set.

Agreement with existing TF-binding data

The *k*-mer-binding profiles and PWMs derived from universal PBM experiments are typically very consistent with TF-binding data obtained by other *in vitro* approaches. Databases such as TRANSFAC⁵⁶ and JASPAR⁵⁷ contain hundreds of matrices constructed from existing binding data. (TRANSFAC tends to be more inclusive, whereas JASPAR is manually curated and limited to a smaller number of TFs with high-confidence data.) Our PBM data nearly always agree with the corresponding entries in these databases at a coarse level, especially JASPAR. Slight discrepancies are not surprising, especially given that the database entries often exhibit ascertainment bias reflecting which particular sequences were chosen to be examined by the investigators. Furthermore, single PWMs in TRANSFAC are frequently derived from binding sequence data compiled from multiple experimental methods. In contrast, universal PBMs provide a uniform, unbiased platform for identifying comprehensive TF-binding profiles. Large discrepancies between PBMs and existing data may also occasionally be observed, but this is also not surprising, given that data in TRANSFAC and JASPAR for identical proteins are not always in agreement with each other¹⁷. This illustrates that motifs in databases and the literature cannot all be taken as a gold standard. Furthermore, even though PBM data do agree with existing binding data, the PBM data provide a richness and level of detail absent from these databases, which typically contain only a handful of sequences.

Comparisons can also be made with *in vivo* binding data generated by alternate methods such as ChIP-chip⁸. There are many reasons why *in vitro* PBM data might not agree with established *in vivo* binding sites, several of which are discussed in INTRODUCTION. TFs may require specific cofactors or post-translational modifications for optimal DNA binding. Furthermore, ligand-binding, heterodimeric protein interactions and associations with other proteins *in vivo* can modulate the binding specificity of a TF through structural changes⁴⁰. Nevertheless, we have observed data from our own PBM experiments to be very consistent with sites known to be bound *in vivo*^{8,17}.

Binding-site representation: *k*-mers versus PWMs

The analysis method described here produces two distinct representations of the DNA-binding specificity of a TF: an exhaustive table of the relative preferences for all *k*-mers, and a mononucleotide PWM (**Fig. 5**). Each representation carries its own set of advantages, and each is suitable for a variety of applications.

The ability to generate a comprehensive list of the relative preferences of a TF for all possible k -mers is one of the most important features of universal PBMs. This offers the opportunity to examine the full landscape of TF binding, including moderate- and low-affinity sequences. Additionally, it provides a high-resolution picture of protein–DNA interactions by conveying information about nucleotide interdependencies. Independent measurements of DNA-binding affinity constants are consistent with k -mer median signal intensities and E -scores derived from PBMs, including for TFs and k -mers exhibiting nucleotide interdependence⁹. Complete k -mer-binding profiles also enable the detailed comparison of the binding specificities of structurally similar TFs that otherwise share the same overall motif. For example, **Figure 9** shows a comparison of the 8-mer E -scores for two related mouse TFs, Lhx2 and Lhx4. Although these TFs exhibit identical motifs and bind the same highest affinity 8-mers, they differ significantly in their preferred lower-affinity binding sites¹⁷.

Nevertheless, PWMs have proven to be a reliable, useful method for binding-site representation. In their compactness, they present a much more intuitive picture of the binding specificity of a TF than a lengthy list of individual k -mer scores. For TFs that make extensive contacts with DNA, the PWMs derived from universal PBMs are particularly useful because they can be substantially longer than 8 base pairs, owing to the incorporation of information from many gapped k -mer patterns. (By considering different gapped patterns as candidate seeds, the resulting PWM will be anchored on the eight most informative positions within the motif.) Finally, most existing softwares for searching genomic occurrences of TF-binding sites are designed to take PWMs as input¹². Such analyses enable the prediction of direct regulatory targets of individual TFs in relatively compact eukaryotic genomes, such as yeast. In higher eukaryotes, where TFs often bind at a much greater distance from their target genes, more complicated prediction strategies are necessary^{58,59}.

We expect that the use of k -mer-binding data, rather than PWMs, for searching genomic sequence will enable more accurate prediction of TF-binding sites across the genome. Traditionally, PWMs have been used when only limited experimental binding data existed for a particular TF, allowing the preferences of the TF for all other sequences to be approximated. Now, universal PBMs allow the generation of comprehensive binding data for all k -mers. This constitutes a significant paradigm shift in the study of gene regulation. Consequently, new methodologies will be needed to score candidate regulatory regions of genomes according to the relative preferences of TFs over all possible k -mers. New databases to store these extensive k -mer-specific data will be necessary; the recently developed UniPROBE database hosts both k -mer-specific data and PWMs for published universal PBM data⁶⁰. We expect universal PBMs to provide valuable data sets for understanding the regulatory processes that govern gene expression in all species.

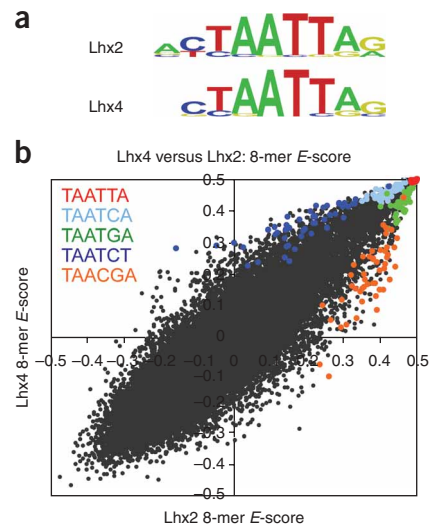


Figure 9 | Differences in k -mer-binding profiles for highly similar TFs. (a) Sequence logos representing nearly identical PWMs derived from PBM experiments for the two mouse homeodomain TFs, Lhx2 and Lhx4. (b) Scatter plot comparing 8-mer scores for these same TFs. 8-mers containing each 6-mer sequence (upper left) are highlighted, revealing clear, systematic differences in the sequence preferences of these TFs for lower-affinity 8-mers despite identical preferences for the same highest-affinity 8-mers (containing TAATTA). This figure has been adapted with permission from ref. 17.

ACKNOWLEDGMENTS We thank Anthony Philippakis for helpful discussion, Andrew Gehrke for technical assistance and Manuel Llinas and Stephen Gisselbrecht for helpful comments and critical reading of the manuscript. M.F.B. and M.L.B. were funded by NIH/NHGRI grant no. R01 HG003985.

COMPETING INTERESTS STATEMENTS The authors declare competing financial interests (see the HTML version of this article for details).

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

1. Ho, S.W., Jona, G., Chen, C.T., Johnston, M. & Snyder, M. Linking DNA-binding proteins to their recognition sequences by using protein microarrays. *Proc. Natl. Acad. Sci. USA* **103**, 9940–9945 (2006).
2. Reece-Hoyes, J.S. *et al.* A compendium of *Caenorhabditis elegans* regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biol.* **6**, R110 (2005).
3. Adryan, B. & Teichmann, S.A. FlyTF: a systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*. *Bioinformatics* **22**, 1532–1533 (2006).

4. Gray, P.A. *et al.* Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science* **306**, 2255–2257 (2004).
5. Messina, D.N., Glasscock, J., Gish, W. & Lovett, M. An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.* **14**, 2041–2047 (2004).
6. Bulyk, M.L., Gentale, E., Lockhart, D.J. & Church, G.M. Quantifying DNA–protein interactions by double-stranded DNA arrays. *Nat. Biotechnol.* **17**, 573–577 (1999).
7. 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).
8. Mukherjee, S. *et al.* Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* **36**, 1331–1339 (2004).
9. Berger, M.F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006).
10. 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).

11. Philippakis, A.A., Qureshi, A., Berger, M.F. & Bulyk, M.L. Design of compact, universal DNA microarrays for protein binding microarray experiments. *J. Comput. Biol.* **15** (2008).
12. Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).
13. Man, T.K. & Stormo, G.D. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.* **29**, 2471–2478 (2001).
14. 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).
15. Benos, P.V., Bulyk, M.L. & Stormo, G.D. Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Res.* **30**, 4442–4451 (2002).
16. McCord, R.P., Berger, M.F., Philippakis, A.A. & Bulyk, M.L. Inferring condition-specific transcription factor function from DNA binding and gene expression data. *Mol. Syst. Biol.* **3**, 100 (2007).
17. Berger, M.F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266–1276 (2008).
18. Fried, M. & Crothers, D.M. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res.* **9**, 6505–6525 (1981).
19. Garner, M.M. & Revzin, A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Res.* **9**, 3047–3060 (1981).
20. Galas, D.J. & Schmitz, A. DNase footprinting: a simple method for the detection of protein–DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).
21. Bowen, B., Steinberg, J., Laemmli, U.K. & Weintraub, H. The detection of DNA-binding proteins by protein blotting. *Nucleic Acids Res.* **8**, 1–20 (1980).
22. Jost, J.P., Munch, O. & Andersson, T. Study of protein–DNA interactions by surface plasmon resonance (real time kinetics). *Nucleic Acids Res.* **19**, 2788 (1991).
23. 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).
24. Linnell, J. *et al.* Quantitative high-throughput analysis of transcription factor binding specificities. *Nucleic Acids Res.* **32**, e44 (2004).
25. Wang, J.K., Li, T.X., Bai, Y.F. & Lu, Z.H. Evaluating the binding affinities of NF- κ B p50 homodimer to the wild-type and single-nucleotide mutant Ig- κ B sites by the unimolecular dsDNA microarray. *Anal. Biochem.* **316**, 192–201 (2003).
26. Warren, C.L. *et al.* Defining the sequence-recognition profile of DNA-binding molecules. *Proc. Natl. Acad. Sci. USA* **103**, 867–872 (2006).
27. Shumaker-Parry, J.S., Aebersold, R. & Campbell, C.T. Parallel, quantitative measurement of protein binding to a 120-element double-stranded DNA array in real time using surface plasmon resonance microscopy. *Anal. Chem.* **76**, 2071–2082 (2004).
28. Maerkl, S.J. & Quake, S.R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).
29. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
30. Reid, J.L., Iyer, V.R., Brown, P.O. & Struhl, K. Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Mol. Cell* **6**, 1297–1307 (2000).
31. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
32. Bulyk, M.L. DNA microarray technologies for measuring protein–DNA interactions. *Curr. Opin. Biotechnol.* **17**, 422–430 (2006).
33. van Steensel, B., Delrow, J. & Henikoff, S. Chromatin profiling using targeted DNA adenine methyltransferase. *Nat. Genet.* **27**, 304–308 (2001).
34. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316**, 1497–1502 (2007).
35. Wei, C.L. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124**, 207–219 (2006).
36. Wold, B. & Myers, R.M. Sequence census methods for functional genomics. *Nat. Methods* **5**, 19–21 (2008).
37. Pompeani, A.J. *et al.* The *Vibrio harveyi* master quorum-sensing regulator, LuxR, a TetR-type protein is both an activator and a repressor: DNA recognition and binding specificity at target promoters. *Mol. Microbiol.* Aug 14 [Epub] (2008).
38. De Silva, E.K. *et al.* Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc. Natl. Acad. Sci. USA* **105**, 8393–8398 (2008).
39. Choi, Y. *et al.* Microarray analyses of newborn mouse ovaries lacking Nobox. *Biol. Reprod.* **77**, 312–319 (2007).
40. Marmorstein, R. & Fitzgerald, M.X. Modulation of DNA-binding domains for sequence-specific DNA recognition. *Gene* **304**, 1–12 (2003).
41. Benos, P.V., Lapedes, A.S. & Stormo, G.D. Is there a code for protein–DNA recognition? Probabilistic. *Bioessays* **24**, 466–475 (2002).
42. Blancafort, P., Segal, D.J. & Barbas, C.F. III Designing transcription factor architectures for drug discovery. *Mol. Pharmacol.* **66**, 1361–1371 (2004).
43. Gommans, W.M., Haisma, H.J. & Rots, M.G. Engineering zinc finger protein transcription factors: the therapeutic relevance of switching endogenous gene expression on or off at command. *J. Mol. Biol.* **354**, 507–519 (2005).
44. Wilson, D.S. & Desplan, C. Structural basis of Hox specificity. *Nat. Struct. Biol.* **6**, 297–300 (1999).
45. Joshi, R. *et al.* Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* **131**, 530–543 (2007).
46. Chan, S.K., Popperl, H., Krumlauf, R. & Mann, R.S. An extradenticle-induced conformational change in a HOX protein overcomes an inhibitory function of the conserved hexapeptide motif. *EMBO J.* **15**, 2476–2487 (1996).
47. Walhout, A.J. *et al.* GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol.* **328**, 575–592 (2000).
48. 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).
49. GuhaThakurta, D. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res.* **34**, 3585–3598 (2006).
50. Chen, X., Hughes, T.R. & Morris, Q. RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. *Bioinformatics* **23**, i72–i79 (2007).
51. Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.* **16**, 962–972 (2006).
52. Foat, B.C., Morozov, A.V. & Bussemaker, H.J. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**, e141–e149 (2006).
53. 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).
54. Workman, C.T. *et al.* enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res.* **33**, W389–W392 (2005).
55. Huber, B.R. & Bulyk, M.L. Meta-analysis discovery of tissue-specific DNA sequence motifs from mammalian gene expression data. *BMC Bioinformatics* **7**, 229 (2006).
56. Wingender, E. *et al.* TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* **28**, 316–319 (2000).
57. 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).
58. Warner, J.B. *et al.* Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat. Methods* **5**, 347–353 (2008).
59. Pennacchio, L.A. & Rubin, E.M. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* **2**, 100–109 (2001).
60. Newburger, D. & Bulyk, M.L. UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acids Res.* **37**, D77–D82 (2009).