nature biotechnology

De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis

Polly M Fordyce[1,2,6], Doron Gerber[3,6], Danh Tran[4], Jiashun Zheng[1], Hao Li[1,5], Joseph L DeRisi[1,2,6] & Stephen R Quake[2,4,6]

Gene expression is regulated in part by protein transcription factors that bind target regulatory DNA sequences. Predicting DNA binding sites and affinities from transcription factor sequence or structure is difficult; therefore, experimental data are required to link transcription factors to target sequences. We present a microfluidics-based approach for *de novo* discovery and quantitative biophysical characterization of DNA target sequences. We validated our technique by measuring sequence preferences for 28 *Saccharomyces cerevisiae* transcription factors with a variety of DNA-binding domains, including several that have proven difficult to study by other techniques. For each transcription factor, we measured relative binding affinities to oligonucleotides covering all possible 8-bp DNA sequences to create a comprehensive map of sequence preferences; for four transcription factors, we also determined absolute affinities. We expect that these data and future use of this technique will provide information essential for understanding transcription factor specificity, improving identification of regulatory sites and reconstructing regulatory interactions.

Recent evidence suggests that knowledge of both strongly and weakly bound sequences and their interaction affinities is required for an accurate understanding of transcriptional regulation. Weak-affinity sites are evolutionarily conserved, make significant contributions to overall transcription[1,2] and may allow closely related transcription factors to mediate different transcriptional responses[3]. In addition, quantitative models require both strongly and weakly bound sequences and their binding affinities to recapitulate transcriptional responses[4–7].

Unfortunately, quantitative data detailing transcription factor binding are often lacking, even for model organisms. *In vivo* immunoprecipitation-based methods, such as ChIP-chip[8] and ChIP-SEQ[9], provide genome-wide information about promoter occupancy. However, these techniques require knowledge of physiological states under which transcription factors are bound to promoters, cannot distinguish whether a transcription factor contacts DNA directly or is tethered by means of another DNA-binding protein, and do not measure affinities.

In vitro methods complement *in vivo* data by measuring binding affinities, distinguishing whether transcription factors directly bind DNA, and allowing manipulation of post-translational modifications and buffer conditions. Furthermore, *in vitro* methods can be used without knowledge of the conditions under which transcription factors are active. However, current *in vitro* methods cannot simultaneously discover both high- and low-affinity target sequences and measure their affinities. Electromobility shift assays[10], DNAse footprinting[11] and surface plasmon resonance[12] require prior knowledge of potential binding sites, precluding motif discovery. Conversely, selection techniques (e.g., SELEX) and one-hybrid systems[13] discover motifs from a large sequence space, but recover only the most strongly bound sequences, without affinity information. Protein binding microarrays (PBMs)[3,14–18] can discover both strongly and weakly bound sequences but cannot measure reactions at equilibrium, preventing affinity measurements. PBMs also suffer from reduced sensitivity: a recent study using PBMs to probe transcription factor binding in *S. cerevisiae* failed to recover consensus motifs for 49 of 101 transcription factors with previous evidence of direct DNA binding[15]. Embedding immobilized DNA in hydrogels[19] extends the PBM technique to allow affinity and kinetic measurements, but this approach can analyze binding to only ~100 DNA sequences at a time.

An alternative approach is mechanically induced trapping of molecular interactions (MITOMI), a technique that uses a microfluidic device to measure binding interactions at equilibrium, allowing construction of detailed maps of binding energy landscapes. The first-generation MITOMI device measured 640 parallel interactions and required DNA libraries that were specific to a particular transcription factor[20].

Here we report a second-generation MITOMI device (MITOMI 2.0) capable of measuring 4,160 parallel interactions. Devices were fabricated in polydimethylsiloxane (PDMS) using multilayer soft lithography; each device had 4,160 unit cells and ~12,555 valves

[1]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California, USA. [2]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. [3]Department of Life Sciences, Bar Ilan University, Ramat Gan, Israel. [4]Departments of Bioengineering and Applied Physics, Stanford University, Palo Alto, California, USA. [5]Center for Theoretical Biology, Peking University, Beijing, China. [6]These authors contributed equally to this work. Correspondence should be addressed to S.R.Q. (quake@stanford.edu) or J.L.D. (joe@derisilab.ucsf.edu).

Received 17 March; accepted 5 August; published online 29 August 2010; doi:10.1038/nbt.1675

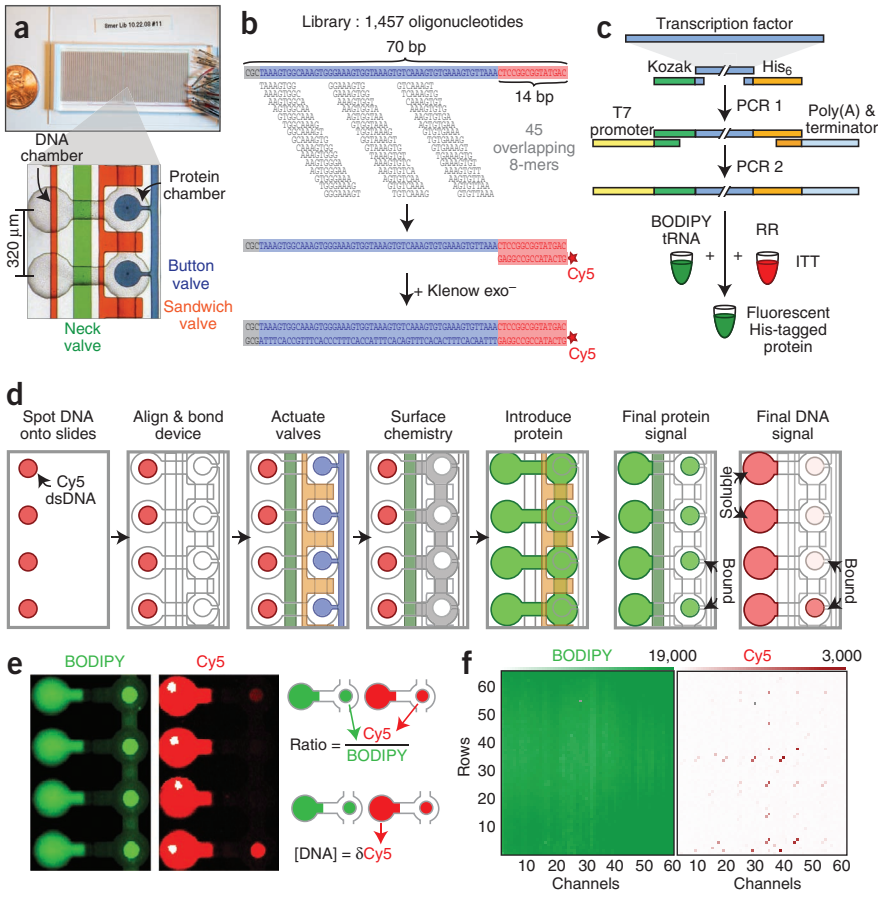

Figure 1 Overall experimental design and procedure. (**a**) Microfluidic device hybridized to glass slide. Unit cells contain two chambers (a 'DNA chamber' and a 'protein' chamber) controlled by three valves: a 'neck' valve (green) separates the two chambers; a 'sandwich' valve (orange) isolates unit cells; and a 'button' valve (blue) protects molecular interactions. (**b**) DNA 8-mer library design. Each 70-bp oligonucleotide contains 45 overlapping 8-mers, a 3-bp GC-clamp at the 5′ end and an identical 14-bp sequence at the 3′ end for Cy5 labeling and primer extension. (**c**) PCR generation of linear templates for protein expression. In PCR1, template-specific primers attach a Kozak sequence, 6× His tag and universal overhangs. In PCR2, universal primers add a T7 promoter, poly-A tail and T7 terminator. *In vitro* transcription and translation (ITT) of this template in rabbit reticulocyte lysate (RR) with BODIPY-labeled, lysine-charged tRNA produces labeled, His-tagged protein. (**d**) Overview of experimental procedure. Devices are manually aligned to a spotted microarray. Neck valves are closed to protect DNA within chambers, and slide surfaces are derivatized with anti-pentaHis antibodies below the button (white) and passivated elsewhere (gray). Lysate containing fluorescently labeled His-tagged transcription factors is introduced and neck valves are opened to allow interaction between transcription factors and DNA; sandwich valves are closed to isolate each unit cell. After an incubation, button valves are pressurized to protect protein–DNA interactions, unbound DNA and proteins are washed out, and the device is scanned. δ is a proportionality constant. (**e**) Scanned picture showing final protein (BODIPY, left) and DNA (Cy5, right) intensities in the chamber and under the button. (**f**) Arrays showing example protein intensities (left) and DNA intensities (right) under the button for each unit cell within a device.

to control fluid flow (**Fig. 1a** and **Supplementary Fig. 1**). Each unit cell contained a DNA chamber and a protein chamber, controlled by micromechanical valves—a 'neck' valve, 'sandwich' valves and a 'button' valve (**Fig. 1a**). Unit cells were programmed with particular DNA sequences by aligning and bonding the device with a noncovalently spotted DNA microarray containing a library of 1,457 double-stranded Cy5-labeled oligonucleotides. To accommodate all 65,536 DNA 8-mers, we designed each 70-bp oligonucleotide to contain 45 overlapping, related 8-mer de Bruijn sequences[21] (**Fig. 1b**). Each oligonucleotide sequence appeared in at least two unit cells.

To evaluate the performance of this technique, we measured DNA binding for 28 *S. cerevisiae* transcription factors from ten different families (**Supplementary Table 1**). Of these, there was prior evidence for 26 transcription factors, of direct, sequence-specific DNA binding, and 2 transcription factors had no previously annotated literature motifs, despite multiple previous attempts[14,15,22].

All transcription factor protein was produced by *in vitro* transcription and translation. PCR-generated linear expression templates were added directly to rabbit reticulocyte lysate off-chip in the presence of a small fraction of BODIPY-labeled, lysine-charged tRNA to produce BODIPY-labeled, His-tagged transcription factors (**Fig. 1c** and **Supplementary Fig. 2**). In each experiment, ~50 µl of extract (~100 ng of protein) was loaded into the device.

After alignment to DNA microarrays, slide surfaces within the protein chamber were derivatized with anti-pentaHis antibodies beneath the button valve and passivated elsewhere (**Fig. 1d**). Introduction of His-tagged transcription factors into both chambers solubilized spotted DNA, allowing transcription factors and DNA to interact. Transcription factor–DNA complexes were captured on the surface beneath the button valve during a ~1 h incubation; rapid closure of the button valve trapped interactions at equilibrium concentrations before a final wash to remove unbound material before imaging[20].

BODIPY intensities under the button valve reflect the number of surface-bound protein molecules; Cy5 intensities under the button valve reflect the number of DNA molecules bound by surface-immobilized protein (**Fig. 1d–f**). Therefore, the ratio of Cy5 to BODIPY fluorescence is linearly proportional to the number of protein molecules with bound DNA, or protein fractional occupancy. Cy5 intensities within the DNA chamber reflect the amount of soluble DNA available for binding.

All 28 transcription factors showed oligonucleotide-specific variations in bound Cy5 intensities, demonstrating marked preferences for individual oligonucleotides (**Fig. 2a** and **Supplementary Fig. 3**). By contrast, the distribution of intensities for rabbit reticulocyte extract alone was well fit by a Gaussian distribution (reduced $\chi^2 = 1.0$, $P = 0.47$), establishing that binding is due to expressed transcription factors and not components of the *in vitro* transcription and translation system (**Fig. 2a**).

Variations in fluid flow between channels can lead to differences in the number of protein molecules beneath each button valve. To account for these differences and generate a quantity proportional

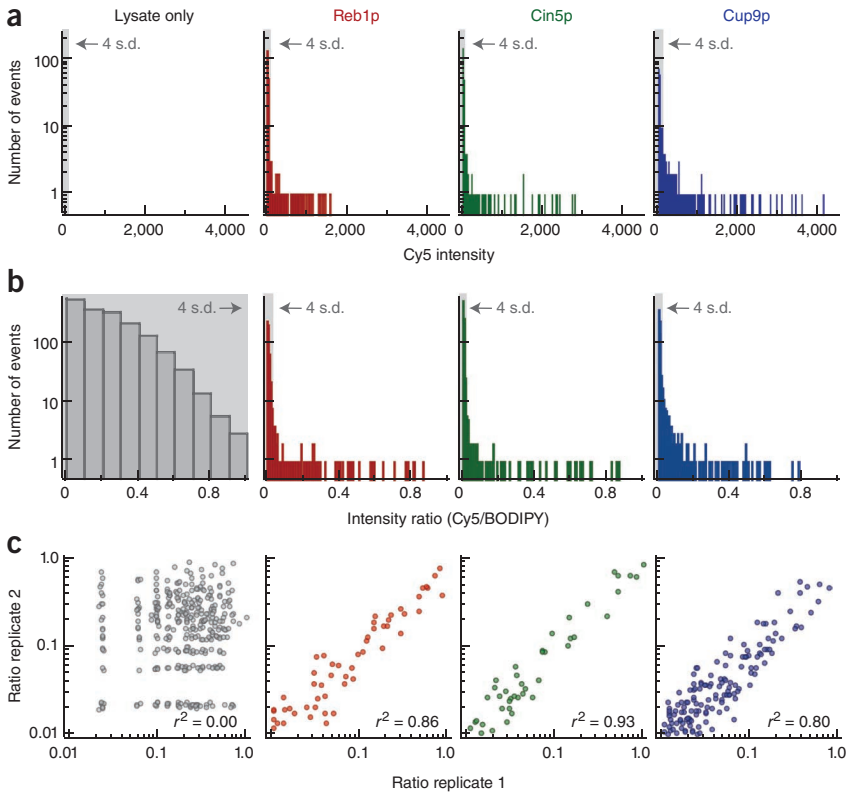

Figure 2 Detailed analysis of measured Cy5 intensities and fluorescence intensity ratios (Cy5/BODIPY-FL) for rabbit reticulocyte lysate alone, Reb1p, Cin5p and Cup9p. (**a**) Distribution of measured Cy5 intensities for all oligonucleotides. Light gray box indicates measurements within 4 s.d. of the mean (as determined by a Gaussian fit). Measured Cy5 intensities for rabbit reticulocyte lysate alone are well fit by a Gaussian distribution (reduced $\chi^2 = 1.0$, $P = 0.47$). For all transcription factors, measured Cy5 intensities deviate significantly from a Gaussian distribution, with measured events many s.d. above the mean. (**b**) Distribution of measured intensity ratios for all oligonucleotides. Light gray box indicates measurements within 4 s.d. of the mean (as determined by a Gaussian fit). Measured intensity ratios in the presence of transcription factors deviate significantly from a normal distribution (**Supplementary Table 2**). (**c**) Correlation between ratios measured for the same oligonucleotide at two separate locations within the device.

to fractional occupancy, Cy5 intensities were normalized by BODIPY intensities to yield a dimensionless intensity ratio (Cy5 intensity/BODIPY intensity) (**Fig. 1e**). Intensity ratios also showed strong preferences for individual oligonucleotide sequences, with no clear preference detected for rabbit reticulocyte lysate alone (**Fig. 2b**, **Supplementary Fig. 4** and **Supplementary Table 2**). Intensity ratios were well correlated both between measurements of the same 70-mer oligonucleotide at different locations within a given device (**Fig. 2c** and **Supplementary Table 3**) and between experiments (**Supplementary Fig. 5**).

Binding affinity can be described by a single-site binding model relating intensity ratio (r) to DNA concentration ($[D]$); K_d, the DNA concentration at which measured intensities reach half their maximum value (r_{max}) provides a quantitative measure of binding affinity.

$$r = \frac{r_{max} \cdot [D]}{[D] + K_d} \quad (1)$$

At low DNA concentrations, measured intensity ratios are approximately inversely proportional to K_d. Calibrated measurements of DNA chamber intensities in our experiments establish that soluble DNA concentrations are indeed low (150 ± 25 nM, mean ± s.e.m.) (**Supplementary Fig. 6**), suggesting it might be possible to accurately estimate interaction affinities from intensity ratios measured at a single, low DNA concentration.

To test this hypothesis, we first measured concentration-dependent binding for four transcription factors (Cbf1p, Cin5p, Pho4p and Yap1p) from two different families, each interacting with ten oligonucleotides from the 8-mer DNA library. We then globally fit equation (1) over all oligonucleotides at all concentrations to get accurate K_d measurements (**Fig. 3a** and **Supplementary Figs. 7–9**).

Next, we calculated K_d values for the exact same oligonucleotides from single-concentration measurements. The low DNA concentration used for these measurements prevented direct determination of r_{max}, a parameter that depends on quantities that vary between experiments (e.g., amount and intensity of BODIPY and Cy5 dyes incorporated during protein and DNA library production, respectively), and must be empirically determined. K_d values from concentration-dependent binding can be used to 'calibrate' the appropriate r_{max} value (**Supplementary Methods** and **Supplementary Tables 4** and **5**). Single-concentration K_d values calculated using calibrated r_{max} values were in excellent agreement with those derived from concentration-dependent binding ($r^2 = 0.90$, $P = 2.1 \times 10^{-19}$) (**Fig. 3b**). Furthermore, once calibrated, r_{max} values can be used to calculate K_d values for all oligonucleotides with signals above background, providing absolute affinities for all 1,457 oligonucleotides with only a few additional measurements (**Fig. 3c**,**d** and **Supplementary Fig. 10**). The range of K_d values calculated here for Pho4p and Cbf1p agree with those measured in previous studies (~10 nM–10 μM)[20], validating our approach. Relative differences in binding affinities between oligonucleotides (the Gibbs free energy upon binding, ΔΔG) can also be calculated using these calibrated r_{max} values (**Supplementary Fig. 11**).

Even in the absence of additional information to calibrate r_{max} values, however, measured intensity ratios provide accurate information about binding affinity. To demonstrate this, we assumed an r_{max} value of 1 for all transcription factors and again compared measured and calculated K_d values. K_d measurements were well correlated ($r^2 = 0.67$, $P = 1.8 \times 10^{-10}$), although individual curves were systematically offset (**Supplementary Fig. 12a**). ΔΔG describes relative affinity differences between oligonucleotides and is therefore less sensitive to these offsets, with stronger correlations ($r^2 = 0.76$, $P = 8.0 \times 10^{-13}$) (**Supplementary Fig. 12b**).

Measured intensity ratios reflect interaction affinities between a given transcription factor and a 70-bp oligonucleotide. Identifying transcription factor target sites requires determination of the precise subsequences responsible for transcription factor binding within each oligonucleotide. Traditionally, analysis of transcription factor binding requires designation of sequences into bound and unbound populations, followed by a search for sequences overrepresented in the bound population, which ignores relative strengths of binding interactions,

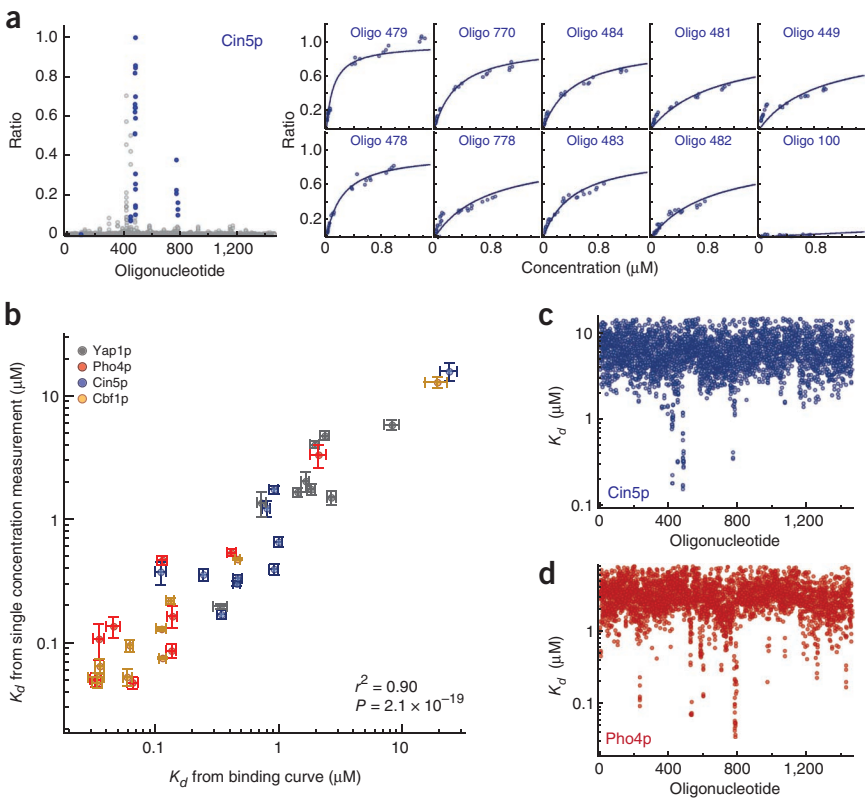

Figure 3 Comparison between K_d values derived from direct measurements of concentration-dependent binding and K_d values calculated from ratio measurements at a single concentration. (**a**) Cin5p measurements. Measured ratio signals for all oligonucleotides (gray) and selected oligonucleotides (blue) (left); concentration-dependent binding for selected oligonucleotides fit to a single-site binding model (right). (**b**) K_d calculated from single-concentration measurements compared with K_d derived from fits concentration-dependent binding for Cin5p (blue), Pho4p (red), Yap1p (gray) and Cbf1p (gold). (**c**) Calculated K_d values for all oligonucleotides for Cin5p. (**d**) Calculated K_d values for all oligonucleotides for Pho4p.

and can be sensitive to the precise threshold used to delineate populations. Here we used a pipeline that incorporates all intensity information for all oligonucleotides to generate a position-specific affinity matrix (PSAM)[23] describing the change in binding affinity upon mutation of a specific position within a consensus sequence (**Supplementary Fig. 13**). Notably, PSAMs describe actual binding affinities for any combination of nucleotides and can be used to calculate predicted affinities to arbitrary sequences.

First, we analyzed all measured intensity ratios using fREDUCE, an enumerative algorithm that searches for sequences whose occurrence within oligonucleotides correlates strongly with their measured signal[24]. For all 28 proteins, fREDUCE returned sequences with strong correlations (**Supplementary Table 6** and **Supplementary Fig. 14**).

Next, the highest-correlated 7- and 8-bp fREDUCE sequences were converted to PSAMs using MatrixREDUCE[23], an algorithm that fits all measured intensity ratios with a statistical mechanical model assessing the effects of individual base-pair substitutions on binding affinity. Because investigations of MatrixREDUCE performance have recommended the use of initial seed sequences derived from enumerative analysis to ensure optimization of global minima[24], the fREDUCE sequences were used as seeds. MatrixREDUCE assumes that the free energy contributions of each position in the binding site are independent; although this is known to be false in some instances, we use linear motifs here to compare our results with the largest possible set of previous literature.

To choose the single PSAM that best explains measured binding, we compared occupancies predicted by each PSAM for all oligonucleotides in the DNA library with measured intensity ratios (**Supplementary Fig. 15**). Predicted and measured values were well-correlated for almost all transcription factors (**Supplementary Table 7**). For all 26 transcription factors with described motifs, the final recovered motif was in agreement with those previously reported in the literature (**Fig. 4**)[14,15,22]. We also derived PSAMs for two transcription factors that were previously resistant to characterization, Msn1p and Nrg2p, establishing considerably enhanced sensitivity over both ChIP-based and PBM techniques.

Two well-characterized basic helix-loop-helix proteins (Pho4p and Cbf1p) provide a test of the ability to detect both high- and low-affinity target sequences. Pho4p binds both high-affinity (5′-CACGTG-3′) and low-affinity (5′-CACGTT-3′) sites[25]; Cbf1p binds to a degenerate 5′-RTCACRTG-3′ motif[20,26]. For both proteins, we recovered the expected motif variants (**Fig. 4** and **Supplementary Fig. 15**).

Detailed analysis of differences between measured and calculated binding profiles can provide additional information about binding preferences. For example, oligonucleotides with high measured intensity ratios but low predicted occupancies could indicate binding to additional motifs. In addition, this comparison allows investigation of whether free energy contributions at each position within the sequence are truly independent.

For most transcription factors, optimized PSAMs successfully described gross binding properties (e.g., Pho4p, Cin5p, Msn2p and Sko1p; **Supplementary Fig. 16**), albeit with outliers at weak binding energies that may represent cooperative interactions between base-pair substitutions. For a few transcription factors (Rpn4p, Cup9p, Cad1p, Matα2p and Pdr3p), correlations between measured and predicted binding were much weaker ($r^2 < 0.25$). To determine if low correlations resulted from binding to additional target sequences, we used BioPROSPECTOR[27], MDScan[27], MEME[28] and WEEDER[29] to scan for overrepresented sequences within oligonucleotides with high measured intensity ratios (Z-score > 25 for Rpn4p or 75 for Cup9p) but low predicted occupancies (Z-score < 3).

For Rpn4p, although both PBM studies and our initial analysis identified binding to a 5′-GCCACC-3′ motif, ChIP and expression data suggest a T-rich 5′ extension of this motif upstream of Rpn4p target genes. Notably, analysis of the 13 oligonucleotides with discordant measured and predicted binding returned this precise extension, establishing that unexpected binding data can yield biologically relevant results (**Supplementary Fig. 17**).

The Cup9p-optimized PSAM also agreed with previous PBM[15] results (**Fig. 4**); however, 14 sequences showed stronger-than-predicted binding (**Supplementary Fig. 18**). Analysis of these sequences yielded

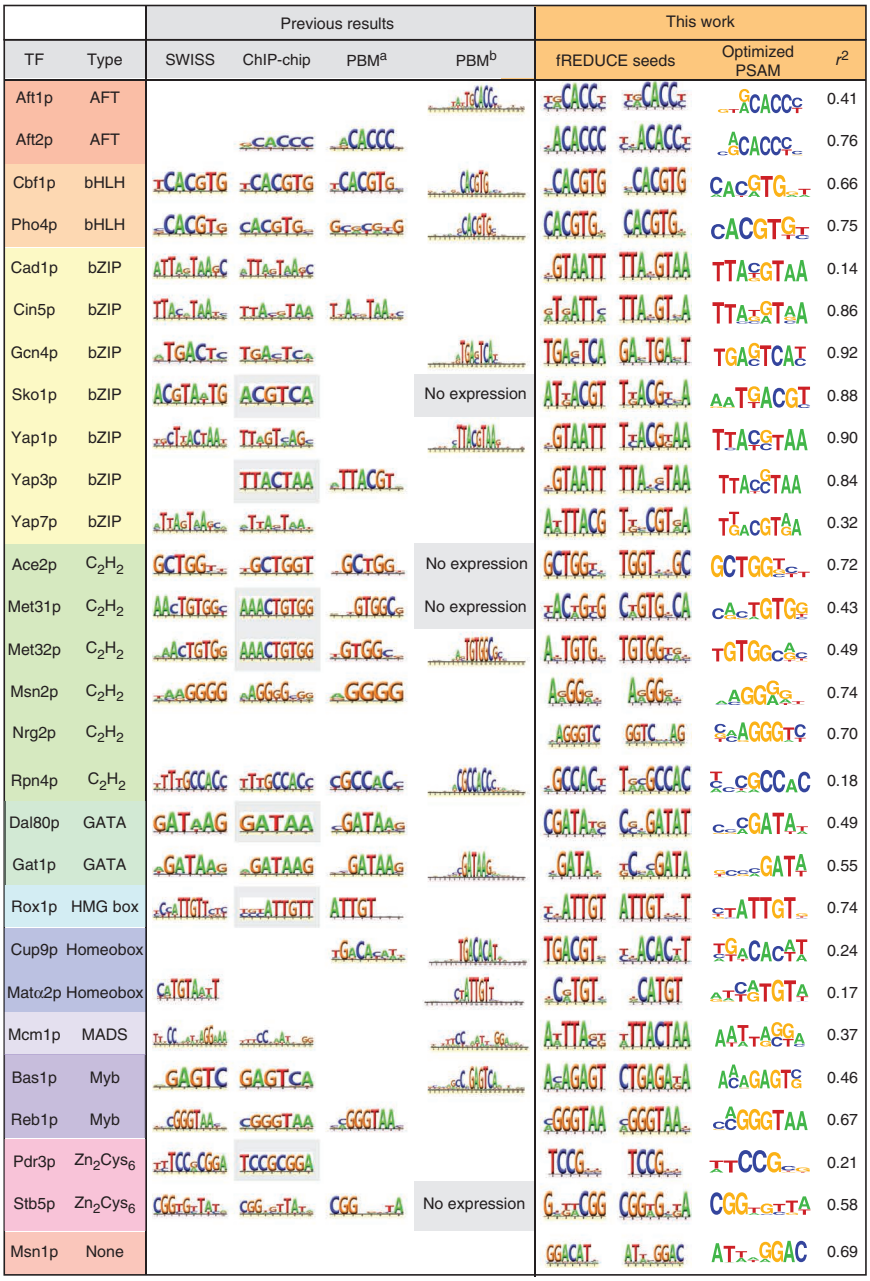

Figure 4 Comparison between motifs found for all 28 *S. cerevisiae* transcription factors and previous literature results (SWISS, SwissRegulon[30]; ChIP-chip, Harbison library[22]; PBM[1], protein binding microarray[14]; PBM[2], protein binding microarray[15]). For ChIP-chip data, boxes shaded in gray represent literature-derived motifs. For PBM[2] results, white boxes represent proteins applied to arrays that did not yield motifs; boxes shaded in gray represent proteins that were not expressed sufficiently to be applied to arrays. fREDUCE Seeds: 7- and 8-bp fREDUCE motifs that correlate most strongly with measured intensities; Optimized PSAM: MatrixREDUCE PSAM represented as an AffinityLogo; r^2: Pearson correlation coefficient between all measured ratio values and protein occupancies predicted by the optimized PSAM.

motifs similar to the optimized PSAM, but with an 'ACGT' core (**Supplementary Fig. 18**, gray box). To assess the affinity of Cup9p for this candidate alternate motif, we measured concentration-dependent binding of Cup9p to the primary motif, candidate secondary motif and several related motifs (**Supplementary Fig. 19a**). A random 2-bp substitution abolished binding, but mutating these bases or the entire second half of the motif to the candidate secondary motif reduced affinity only ~20-fold (**Supplementary Fig. 19b**), confirming weak-affinity binding. Interestingly, this motif is found only 29 times in the genome outside of coding regions, primarily at the boundary of subtelomeric repeats and upstream of genes regulated by iron depletion, metal toxicity or oxidative stress (**Supplementary Table 8**). Although the physiological role of these putative binding sites is unknown, these results demonstrate the ability of MITOMI 2.0 to detect weak but potentially biologically relevant transcription factor binding sites.

For the remaining three transcription factors (Cad1p, Matα2p and Pdr3p), low correlations between predicted and measured binding likely resulted from experimental variability and not binding to additional motifs. Correlations between technical replicates across the device were relatively low (**Supplementary Table 3**), owing to either binding to a limited number of oligonucleotides (Cad1p, **Supplementary Fig. 3**) or large variations in protein coverage (for Matα2p and Pdr3p). Consistent with this, these transcription factors do not bind any oligonucleotides with stronger-than-expected affinity.

The data presented here demonstrate increased sensitivity over current state-of-the-art techniques, detecting sequence-specific binding for several proteins that have failed to yield results in multiple experiments (Cad1p, Msn1p, Nrg2p, Sko1p, Yap7p and Pdr3p). Moreover, these data represent the most comprehensive investigation of biophysical binding affinities to date, including ΔΔG values for 28 transcription factors and K_d values for four transcription factors from two different families (Cbf1p, Cin5p, Pho4p and Yap1p) binding to 1,457 individual sequences. These data can be used to test basic assumptions underlying current models of transcription factor–DNA specificity and to more accurately model cooperativity between nucleotide-binding sites ('nonadditivity').

The DNA library used here is not organism-specific, making this technique useful for a wide range of organisms, including higher eukaryotes and pathogens. In addition, the programmable nature of MITOMI 2.0 allows subsequent detailed examination of unexpected binding phenomena or systematic mutational analysis of candidate motifs through direct observations of concentration-dependent binding. Although these experiments probed transcription factor binding to double-stranded DNA, MITOMI 2.0 can be used, with only minimal changes, to investigate single-stranded DNA binding and RNA binding. When paired with advances in rapid whole-genome sequencing, we anticipate that MITOMI 2.0 characterization of all recognizable transcription factors in a

given proteome will allow transcriptional networks and regulons to be quickly identified and ultimately modeled.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession codes. Gene Expression Omnibus: GPL10817.

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

ACKNOWLEDGMENTS

P.M.F. was supported by a Howard Hughes Medical Institute/Helen Hay Whitney Foundation Postdoctoral Fellowship. J.L.D., S.R.Q. and this work were supported by the Howard Hughes Medical Institute. We thank A. Potanina for assistance with fabrication of microfluidic devices, O. Homann for implementation of PSAM functionality with MochiView and D. Breslow, F. Caro, S. Churchman, M. Dimon, T. Kiers, A. Kistler, C. Nelson, K. Sorber, E. Yeh and I. Zuleta for careful reading of the manuscript.

AUTHOR CONTRIBUTIONS

P.M.F. designed experiments, designed, created and printed the DNA library, made linear expression templates, fabricated microfluidic devices, performed microfluidic experiments assessing concentration-dependent binding and binding to the 8-mer library, analyzed data and wrote the manuscript. D.G. designed experiments, designed and fabricated microfluidic devices and performed microfluidic experiments assessing binding to the 8-mer library. D.T. fabricated microfluidic devices and performed microfluidic experiments assessing binding to the 8-mer library. J.Z. and H.L. analyzed data. S.R.Q. designed experiments, analyzed data and wrote the manuscript. J.L.D. designed experiments, assisted with printing the DNA library, analyzed data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

1. Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.* **16**, 962–972 (2006).
2. Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* **451**, 535–540 (2008).
3. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
4. Kim, H.D. & O'Shea, E.K. A quantitative model of transcription factor-activated gene expression. *Nat. Struct. Mol. Biol.* **15**, 1192–1198 (2008).
5. Segal, E. & Widom, J. From DNA sequence to transcriptional behavior: a quantitative approach. *Nat. Rev. Genet.* **10**, 443–456 (2009).
6. Gertz, J., Siggia, E.D. & Cohen, B.A. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* **457**, 215–218 (2009).
7. Yuh, C.H., Bolouri, H. & Davidson, E.H. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development* **128**, 617–629 (2001).
8. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
9. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
10. 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).
11. 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).
12. 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).
13. Meng, X., Brodsky, M.H. & Wolfe, S.A. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotechnol.* **23**, 988–994 (2005).
14. Badis, G. *et al.* A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell* **32**, 878–887 (2008).
15. Zhu, C. *et al.* High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.* **19**, 556–566 (2009).
16. Berger, M.F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006).
17. Berger, M. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266–1276 (2008).
18. De Silva, E.K. *et al.* Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc. Natl. Acad. Sci. USA* **105**, 8393–8398 (2008).
19. Bonham, A.J., Neumann, T., Tirrell, M. & Reich, N.O. Tracking transcription factor complexes on DNA using total internal reflectance fluorescence protein binding microarrays. *Nucleic Acids Res.* **37**, 94 (2009).
20. Maerkl, S.J. & Quake, S.R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).
21. Ralston, A. De Bruijn sequences-a model example of the interaction of discrete mathematics and computer science. *Math. Mag.* **55**, 131–143 (1982).
22. Harbison, C.T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
23. 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).
24. Wu, R., Chaivorapol, C., Zheng, J., Li, H. & Liang, S. fREDUCE: detection of degenerate regulatory elements using correlation with expression. *BMC Bioinformatics* **8**, 399 (2007).
25. Vogel, K., Horz, W. & Hinnen, A. The two positively acting regulatory proteins PHO2 and PHO4 physically interact with PHO5 upstream activation regions. *Mol. Cell. Biol.* **9**, 2050–2057 (1989).
26. Wieland, G. *et al.* Determination of the binding constants of the centromere protein Cbf1 to all 16 centromere DNAs of *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **29**, 1054–1060 (2001).
27. Liu, Y. *et al.* A suite of web-based programs to search for transcriptional regulatory motifs. *Nucleic Acids Res.* **32**, W204–W207 (2004).
28. Bailey, T.L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
29. Pavesi, G. *et al.* MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Res.* **34**, W566–W570 (2006).
30. Pachkov, M., Erb, I., Molina, N. & Van Nimwegen, E. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res.* **35**, D127–D131 (2007).

ONLINE METHODS

Oligonucleotide sequence files and data for all transcription factors are available for download at http://derisilab.ucsf.edu.

DNA library and transcription-factor production. All possible 65,536 8-bp DNA sequences were assembled into a maximally compact de Bruijn sequence that was subsequently divided over 1,457 oligonucleotides. Sequences were hybridized to a Cy5-labeled oligonucleotide and extended using Klenow fragment (exo-) (New England Biolabs) to produce Cy5-labeled dsDNA. Cy5-labeled dsDNA was diluted to a final concentration of 1.25 μM in 3× SSC with polyethylene glycol (PEG) (Fluka) and D-(+)-trehalose dihydrate (Fluka) (for enhanced subsequent solubility) and printed onto custom 2″ × 3″ ThermoFisher Scientific SuperChip Epoxysilane slides (ThermoFisher Scientific) using a DeRisi lab custom microarrayer.

A two-step PCR reaction was used to amplify transcription factor coding sequences and add appropriate upstream and downstream sequences for efficient transcription and translation in rabbit reticulocyte lysate (Promega) (**Supplementary Fig. 2**).

Microfluidic device fabrication and experimental procedure. Flow and control molds were fabricated on 4″ silicon wafers using positive (SPR 220-7.0) and negative (SU-8) photoresists, respectively. PDMS devices were produced and the MITOMI experimental procedure was performed as described previously[20].

Initial data analysis and normalization. Median Cy5 and BODIPY fluorescence intensities varied somewhat between experiments. To facilitate comparisons between transcription factors, Cy5 intensity distributions were fit to a Gaussian distribution and this Gaussian mean was subtracted from all measurements to center the background distribution around zero. Fluorescence intensity ratios were calculated by dividing Cy5 fluorescence intensities by BODIPY fluorescence intensities; ratios were similarly normalized such that the background was centered around zero, and further normalized such that the maximum measured intensity was 1.

Motif finding pipeline. We searched for 7- and 8-bp sequences that correlated most strongly with measured intensity ratios using fREDUCE. Both doubly- (R, Y, S, W, K, M) and triply- (B, D, H, V) degenerate IUPAC bases were included, and both the forward sequence and its reverse complement were analyzed. The most strongly correlated 7-bp and 8-bp sequences were then used as seeds for MatrixREDUCE analysis, with additional unspecified base pairs added to either side of the 7-bp seed to standardize length.

Occupancy profile calculations. We calculated predicted occupancy profiles from PSAMs using a slight modification of the MatrixREDUCE formalism to reflect the fact that, in our assay, transcription factors are surface-immobilized and DNA sequences are in solution (**Supplementary Methods**).

doi:10.1038/nbt.1675