nature genetics

Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays

Sonali Mukherjee[1,2,9], Michael F Berger[1,3,9], Ghil Jona[4], Xun S Wang[1,5], Dale Muzzey[1,3], Michael Snyder[4,6], Richard A Young[5,7] & Martha L Bulyk[1–3,8]

We developed a new DNA microarray-based technology, called protein binding microarrays (PBMs), that allows rapid, high-throughput characterization of the *in vitro* DNA binding–site sequence specificities of transcription factors in a single day. Using PBMs, we identified the DNA binding–site sequence specificities of the yeast transcription factors Abf1, Rap1 and Mig1. Comparison of these proteins' *in vitro* binding sites with their *in vivo* binding sites indicates that PBM-derived sequence specificities can accurately reflect *in vivo* DNA sequence specificities. In addition to previously identified targets, Abf1, Rap1 and Mig1 bound to 107, 90 and 75 putative new target intergenic regions, respectively, many of which were upstream of previously uncharacterized open reading frames. Comparative sequence analysis indicated that many of these newly identified sites are highly conserved across five sequenced *sensu stricto* yeast species and, therefore, are probably functional *in vivo* binding sites that may be used in a condition-specific manner. Similar PBM experiments should be useful in identifying new *cis* regulatory elements and transcriptional regulatory networks in various genomes.

The interactions between transcription factors and their DNA binding sites are an integral part of transcriptional regulatory networks. They control the coordinated expression of thousands of genes during normal growth and in response to external stimuli. Much progress has been made recently in the identification and analysis of mRNA transcript profiles[1,2], locations of *in vivo* binding sites of transcription factors[3–6] and protein-protein interactions[7–10]. But many transcription factors still have unknown DNA binding specificities and regulatory roles.

Earlier technologies aimed at characterizing DNA-protein interactions are time-consuming and not scalable. Microarray-based readout of chromatin immunoprecipitation (ChIP-chip), or genome-wide location analysis, is currently the most widely used high-throughput method for identifying *in vivo* genomic binding sites for transcription factors[3–6]. But some ChIP-chip experiments do not result in significant enrichment of bound fragments in the immunoprecipitated sample. In addition, there may be transcription factors of interest for which a specific antibody is not available or for which the culture conditions or time points that allow its expression and activity are not known.

We previously developed a spotted microarray technology that used primer-extended, double-stranded synthetic DNAs to quantify the differences in binding affinities for various DNA binding–sequence variants. This technology allowed us to distinguish proteins with similar binding-site preferences and to determine the binding specificities of proteins with degenerate sequence preference[11]. Another group recently extended this technology to use surface plasmon resonance[12]. Although surface plasmon resonance can provide kinetic data, it is not currently scalable to a large number of samples.

Here we developed a new *in vitro* DNA microarray technology for genome-scale characterization of the sequence specificities of DNA-protein interactions. This protein-binding microarray (PBM) technology allows the determination of *in vitro* binding specificities of individual transcription factors in a single day, by assaying the sequence-specific binding of those individual transcription factors directly to double-stranded DNA microarrays spotted with a large number of potential DNA-binding sites. A DNA-binding protein of interest is expressed with an epitope tag, purified and then bound directly to a double-stranded DNA microarray. The PBM is then washed to remove any nonspecifically bound protein and labeled with a fluorophore-conjugated antibody specific for the epitope tag (**Fig. 1a**).

We focused our efforts on the genome of the yeast *Saccharomyces cerevisiae* because of its usefulness as a model organism for both experimental and computational studies. Binding-site data from PBMs on yeast transcription factors corresponded well with binding-site specificities determined from ChIP-chip. Moreover, comparative

[1]Division of Genetics, Department of Medicine, Harvard Medical School; [2]Harvard/MIT Division of Health Sciences and Technology, Brigham and Women's Hospital and Harvard Medical School; and [3]Harvard University Graduate Biophysics Program, Harvard Medical School; Boston; Massachusetts 02115; USA. [4]Departments of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA. [5]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [6]Molecular Biophysics and Biochemistry, and Genetics, Yale University, New Haven, Connecticut 06520, USA. [7]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA. [8]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. [9]These authors contributed equally to this work. Correspondence should be addressed to M.L.B. (mlbulyk@receptor.med.harvard.edu).

Published online 14 November 2004; doi:10.1038/ng1473

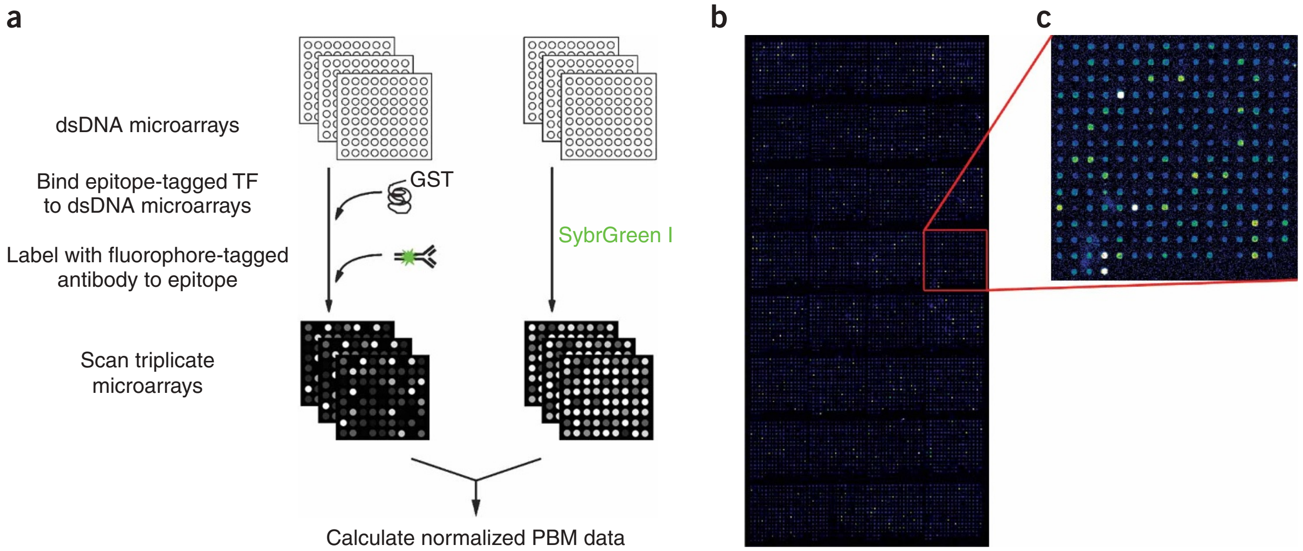


Figure 1 PBM schematic. (**a**) Overview of PBM experiments. (**b**) Whole-genome yeast intergenic microarray bound by Rap1. The fluorescence intensities of the spots are shown in false color, with white indicating saturated signal intensity, red indicating high signal intensity, green indicating moderate signal intensity and blue indicating low signal intensity. (**c**) Magnification of a portion of the whole-genome yeast intergenic microarray bound by Rap1. ds, double-stranded; TF, transcription factor.

sequence analysis of the PBM-derived binding sites indicated that many of the sites bound in PBMs, including some not identified by ChIP-chip, are highly conserved in other *sensu stricto* yeast genomes and therefore are probably functional *in vivo* binding sites that potentially are used in a condition-specific manner. Our PBM technology should aid in the annotation of many regulatory proteins whose DNA-binding specificities have not been characterized and in the construction of gene regulatory networks.

RESULTS

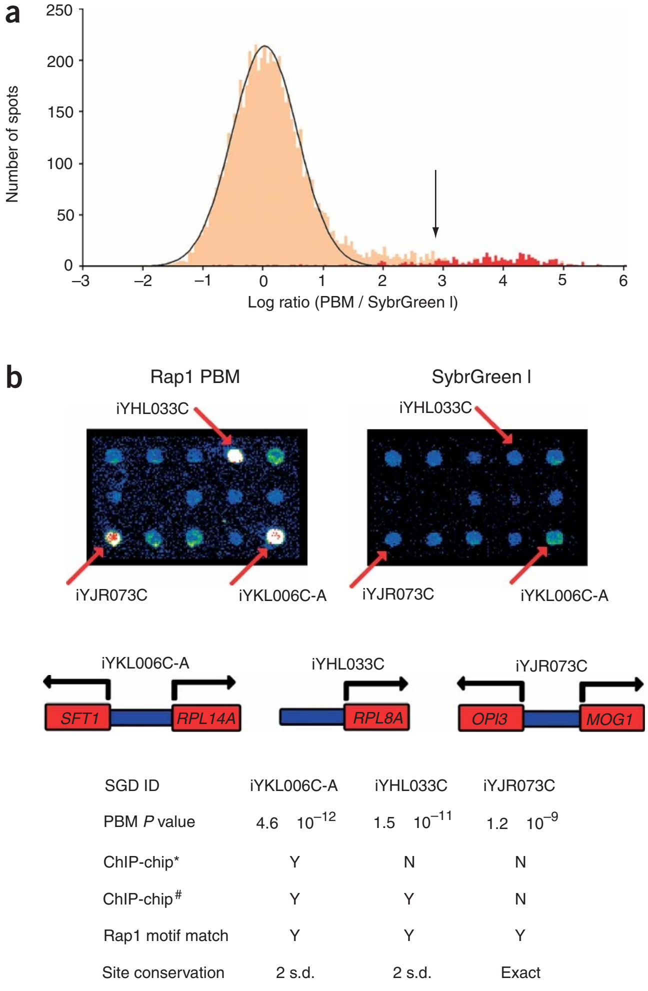
PBM experiments

As a validation of this approach, we bound CBP-FLAG-Rpn4 fusion protein to microarrays spotted with positive and negative control spots for binding by Rpn4. We labeled the protein-bound array with Cy3-conjugated M2 primary antibody to FLAG (Sigma) and scanned it with a microarray scanner (GSI Lumonics ScanArray). Only the spots that contain good matches to the binding-site motif for Rpn4 have high signal intensity (**Supplementary Fig. 1** online). As we previously found that higher signal intensity is generally indicative of higher DNA-protein binding affinity[11], this CBP-FLAG-Rpn4 PBM indicates that our PBM technology is successful in identifying sequence-specific transcription factor binding.

Next, we applied the PBM technology on a genome-wide scale by using whole-genome yeast intergenic arrays in PBM experiments to identify the sequence specificities and target genes of three yeast transcription factors: Abf1, Rap1 and Mig1. Abf1 has a zinc-finger DNA-binding domain, binds origins of replication and regulates ribosome synthesis. Rap1 binds DNA through a Myb-like helix-turn-helix DNA-binding domain and, in addition to regulating ribosome synthesis[13], regulates telomere length and expression at the silent mating-type loci *HML* and *HMR*[14]. Mig1 has a zinc-finger DNA-binding domain and is involved in the repression of glucose-repressed genes[15].

We used Abf1, Rap1 and Mig1, dually tagged at the N terminus with glutathione S-transferase (GST) and His_6, in PBM experiments using microarrays spotted with essentially all the intergenic regions in the yeast genome[3]. The washed, protein-bound microarrays were labeled with Alexa 488-conjugated antibody to GST (Molecular Probes) and scanned with a microarray scanner. The microarray TIF images were quantified using GenePix Pro version 3.0 software. A whole-genome yeast intergenic microarray that was used in a PBM experiment with Rap1 is shown in **Figure 1b**,**c**. Negative control PBMs did not show sequence-specific DNA binding (**Supplementary Fig. 2** online). For each transcription factor, experiments were done in triplicate. We found that the PBM data were highly reproducible, with most spots having a coefficient of variation (i.e., s.d. divided by the mean) <0.3 (**Supplementary Fig. 3** online).

To normalize the PBM data by relative DNA concentration, we stained separate microarrays from the same print run with SybrGreen I (Molecular Probes), which is specific for double-stranded DNA. The distribution of the log ratios of mean PBM to mean SybrGreen I signal intensities for the set of triplicate Rap1 PBM experiments is shown in **Figure 2a**. The spots on the left, whose distribution is fit well by a Gaussian function, are bound nonspecifically by the transcription factor. Conversely, the heavy upper tail of the distribution corresponds to spots that are bound specifically by the transcription factor. For each spot, we calculated a *P* value for specific binding based on the magnitude of its log ratio relative to the standard deviation of the Gaussian distribution. The numbers of unique spots that pass a *P*-value threshold of 0.05, 0.01 or 0.001 for the PBM data of Abf1, Rap1 or Mig1 are shown in **Supplementary Figure 4** online. We used a Bonferroni-corrected *P*-value threshold of 0.001, even though it may increase our false negative rate, to increase the likelihood that spots passing our *P*-value threshold are true positives. This approach disfavors very long intergenic regions, as a single binding site embedded in a long fragment may result in only a moderately high log ratio. Portions of a SybrGreen I–stained microarray and a corresponding Rap1 PBM are shown in **Figure 2b**. The spots with high log ratio PBM data in **Figure 2b** correspond to the intergenic regions directly upstream of known gene targets of Rap1. A complete



SGD ID	iYKL006C-A	iYHL033C	iYJR073C
PBM *P* value	4.6 10^{-12}	1.5 10^{-11}	1.2 10^{-9}
ChIP-chip*	Y	N	N
ChIP-chip#	Y	Y	N
Rap1 motif match	Y	Y	Y
Site conservation	2 s.d.	2 s.d.	Exact

Figure 2 Identifying the specifically bound spots. (**a**) Distribution of ratios of PDM data to SybrGreen I data for Rap1. The arrow indicates those spots passing a *P*-value threshold of 0.001 after correction for multiple hypothesis testing. Indicated in red are spots with an exact match to a sequence belonging to our discovered Rap1 binding-site motif. (**b**) Magnification of intergenic regions, from both PBMs (left) and SybrGreen I–stained microarrays (right), upstream of *RPL14A*, *RPL8A* and *OPI3*, which are known to be direct targets of Rap1. The fluorescence intensities of the spots are shown in false color, color-coded as described for **Figure 1**. PBM *P* values are corrected for multiple hypotheses. Determination of binding in ChIP-chip experiments (Y, yes; N, no) is shown. All regions shown have an exact match to a sequence belonging to the discovered Rap1 motif. For each region, the binding site is conserved across five *sensu stricto* yeast strains, either to within two standard deviations or 100% identical at each position (Exact). The asterisk indicates Rap1 ChIP-chip data from Lee *et al.*[6]; the pound sign (#) indicates Rap1 ChIP-chip data from Lieb *et al.*[5]

listing of *P* values for all intergenic regions for Abf1, Rap1 and Mig1 is available from our website (see URL in Methods). In total, we identified 189, 294 and 79 putative target intergenic regions for Abf1, Rap1 and Mig1, respectively.

Identification of DNA binding site motifs

For each transcription factor, we analyzed the sequences corresponding to spots that had a Bonferroni-corrected *P* value of <0.001 with the motif discovery program BioProspector[16] to identify DNA binding–site motifs for Rap1, Abf1 and Mig1 (**Fig. 3**). Motifs from PBM data passing less stringent *P*-value thresholds are shown in **Supplementary Figure 5** online. The PBM technology allows the identification of both ungapped (*e.g.*, Rap1 and Mig1) and gapped (*e.g.*, Abf1) binding-site motifs. Compared with computational negative controls on matched sets of randomly selected intergenic regions, the group specificity scores[17] derived from the PBM data for each of these three transcription factors were extremely significant (**Fig. 3**). Thus, we have confidence that the PBM data represent true sequence-specific binding of the transcription factors.

The motifs derived from the Rap1, Abf1 and Mig1 PBM data are good matches to the binding-site motifs for these factors derived from the TRANSFAC[18] Professional database (**Fig. 3**). To confirm and further explore the high-resolution binding-site data generated from PBMs, we carried out electrophoretic mobility shift assays (EMSAs). In one example, the bound intergenic region iYPL221W contained a highly significant match (GTGCACGGATTT) to the PBM-derived Rap1 binding-site motif, but it was a poor match to the TRANSFAC Rap1 motif (**Fig. 4a**). The TRANSFAC motif would predict the underlined nucleotides to be unfavorable for Rap1 binding, whereas the PBM motif is somewhat degenerate at these positions. Our EMSA analysis confirmed that Rap1 is capable of high-affinity binding to this sequence (**Fig. 4b**). This is an example of a transcription-factor binding site that would have been missed by using the TRANSFAC motif because of the sparseness and potential ascertainment bias in the TRANSFAC database.

To approximate the potential false-positive rate from PBMs, we determined the fraction of spots passing a *P*-value threshold of 0.001 that was not identified by BioProspector[16] as containing a sequence belonging to the given transcription factor's binding-site motif (**Fig. 2a**). This is by no means a perfect measure; some of these potential false positives could simply have either less-significant matches to the identified motif or multiple occurrences of lower affinity sites that do not belong to the motif. For example, the bound intergenic region iYLL051C, which had only a weak sequence match to the Rap1 PBM motif (**Fig. 4a**), was confirmed by EMSAs to be bound by Rap1 *in vitro* (**Fig. 4b**). This finding suggests that some high-affinity binding sites may not be significant sequence matches to a given transcription factor binding-site motif. Thus, our false-positive rate may be lower than we estimated. Nevertheless, using this approximate measure of false positives, we found that our false-positive rates ranged from ~7% to 9% of 'bound' spots.

Comparison of PBM data and ChIP-chip data

Approximately 6,400, 6,100 and 6,400 unique intergenic PCR products passed our various PBM data quality control filters for Rap1, Abf1 and Mig1, respectively. ChIP-chip data[5,6] were also available for 99.9%, 93.1% and 93.7%, respectively, of these intergenic regions. DNA binding–site motifs identified with the PBM technology for Abf1 and Rap1 corresponded well to motifs determined from analysis of previously published ChIP-chip data passing a *P*-value threshold of 0.001 for these same transcription factors[5,6] (**Fig. 3**). But we could derive the binding site motif for Mig1 from only the PBM data and not from the ChIP-chip data[6]. Unlike Rap1 and Abf1, the intergenic regions identified as bound by Mig1 in PBMs overlapped with only a few regions identified as bound by ChIP-chip (**Fig. 5**). Furthermore, many fewer regions in total were identified as bound by Mig1 in ChIP-chip as compared with PBMs. Because Mig1 is regulated at the level of nuclear localization[15], it is possible that the yeast cultures for the ChIP-chip experiments were such that Mig1 may have been predominantly cytoplasmic. Overall, we identified 107, 90 and 75 putative

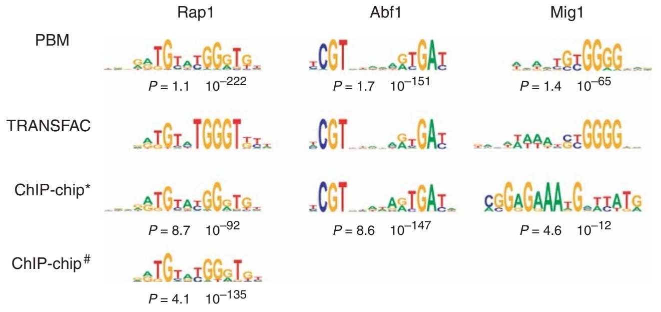


Figure 3 DNA binding site motifs as determined by PBMs compared with motifs derived from ChIP-chip data and from TRANSFAC. Sequence logos were generated essentially as described previously[49]. Group specificity scores are shown. The asterisk indicates Rap1, Abf1 and Mig1 ChIP-chip data from Lee *et al.*[6]; the pound sign indicates Rap1 ChIP-chip data from Lieb *et al.*[5]. Although the Mig1 binding-site motif derived from the ChIP-chip data has a statistically significant group specificity score, it is not a match to either the TRANSFAC or the PBM Mig1 motif. The Pearson correlation coefficients[17] comparing the PBM and ChIP-chip motifs, as well as those comparing each of these motifs versus the motifs present in the TRANSFAC database[18], were as follows: Rap1 PBM versus Lee *et al.*[6] ChIP-chip, 0.992; Rap1 PBM versus Lieb *et al.*[5] ChIP-chip, 0.995; Rap1 PBM versus TRANSFAC, 0.953; Rap1 Lee *et al.*[6] versus Lieb *et al.*[5] ChIP-chip, 0.985; Rap1 Lee *et al.*[6] ChIP-chip versus TRANSFAC, 0.921; Rap1 Lieb *et al.*[5] ChIP-chip versus TRANSFAC, 0.950; Abf1 PBM versus ChIP-chip[6], 0.989; Abf1 PBM versus TRANSFAC, 0.978; Abf1 ChIP-chip[6] versus TRANSFAC, 0.986; Mig1 PBM versus ChIP-chip[6], 0.453; Mig1 PBM versus TRANSFAC, 0.938; Mig1 ChIP-chip[6] versus TRANSFAC, 0.406.

new target intergenic regions for Abf1, Rap1 and Mig1, respectively, including those upstream of 25, 40 and 29 previously uncharacterized open reading frames (ORFs), respectively. (See **Supplementary Fig. 6** online for comparisons using various *P*-value thresholds.)

Sequence conservation of identified binding sites

To find evidence supporting our hypothesis that the regions bound only *in vitro* are probably functional *in vivo* but were not identified previously for some specific biological reason, we mapped the predicted binding sites in *S. cerevisiae* to the orthologous positions in the genomes of *Saccharomyces mikatae*, *Saccharomyces kudriavzevii*, *Saccharomyces bayanus* and *Saccharomyces paradoxus*[19,20]. We found that all amino acid residues important in DNA-protein recognition for Abf1, Rap1 and Mig1 were identical across these five *sensu stricto* species. We examined binding-site conservation in two different ways. First, we considered a site to be conserved if the orthologous sequence in all five species was within two standard deviations of the motif average[21] derived from the set of regions passing a *P*-value threshold of 0.001 in PBMs. Second, we used a strict measure of sequence conservation, requiring 100% sequence identity at all informative nucleotide positions of the transcription-factor binding site in all five species. Although the level of conservation varied for these three transcription factors, the binding sites in regions bound in PBMs were as likely to be conserved as the binding sites in regions bound in ChIP-chip (**Fig. 6**). Furthermore, the regions bound only in PBMs and not in ChIP-chip had approximately the same degree of conservation. PBM experiments identified 23, 70 and 38 putative binding sites for Mig1, Abf1 and Rap1, respectively, that were conserved within two standard deviations in all five species and that were not identified as 'bound' in ChIP-chip experiments[5,6]. Moreover, the regions bound only in PBMs identified between six and ten new conserved sites that are 100% identical across all five species. Given the known conservation level across the *sensu stricto* genomes[20], the probability of observing even a single binding site that is 100% conserved by chance is extremely small. Thus, we believe that the intergenic regions bound in PBMs contain functional *in vivo* binding sites.

Identification of target genes

We examined each set of intergenic regions bound in PBMs to determine whether the candidate target genes, located directly downstream of the bound intergenic regions, were over-represented for particular functional groups[17,22]. A complete listing of all candidate target genes and significantly enriched functional categories for all three transcription factors is provided on our website (see URL in Methods). Of the significantly enriched categories for the candidate target genes of Rap1, a large number are consistent with the known regulatory functions of Rap1 (ref. 13), including the MIPS functional categories for ribosome biogenesis, protein synthesis, structural constituents of the ribosome and cell growth and/or maintenance. In addition, 40 previously uncharacterized ORFs are among the newly identified putative Rap1 target genes. As further evidence of their functional importance, many of the corresponding enriched target genes from the PBM data, including the uncharacterized ORFs *YDR109C*, *YKL151C*, *YIL001W* and *YKL082C*, had upstream Rap1 binding sites that were conserved across all five *sensu stricto* yeast species (**Fig. 6**).

Further characterization of these target genes may identify previously unknown biological functions for Rap1. Ydr109c shows strong homology (BLAST E value = 6.0×10^{-97}) to a number of ribitol kinases, and Ykl151c shows homology (BLAST E value = 2.0×10^{-15}) to a carbohydrate kinase family, suggesting that Rap1 might connect

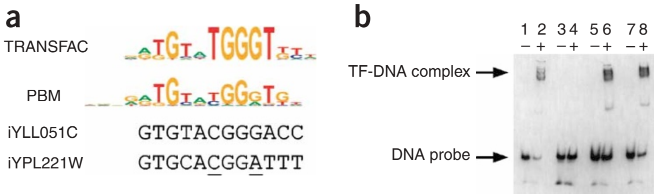


Figure 4 EMSAs of PBM-derived Rap1 binding-site sequences. (**a**) Rap1 binding-site sequences present in the DNA probes corresponding to portions of the intergenic regions iYLL051C ($P = 3.20 \times 10^{-16}$) and iYPL221W ($P = 3.91 \times 10^{-21}$), aligned against the TRANSFAC and PBM-derived Rap1 binding-site sequence logos. (**b**) Lanes 1 and 2, positive control DNA probe; lanes 3 and 4, negative control DNA probe; lanes 5 and 6, DNA probe corresponding to the best Rap1 binding-site sequence that could be identified in the iYLL051C intergenic region; lanes 7 and 8, DNA probe corresponding to the PBM-derived Rap1 binding-site sequence in the iYPL221W intergenic region. The presence (+) or absence (–) of Rap1 protein in the binding reaction is indicated. TF, transcription factor.

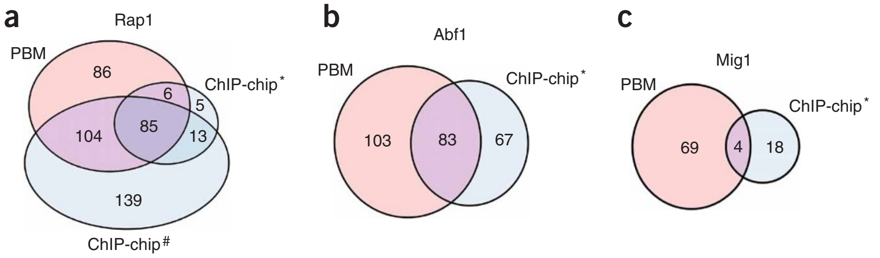


Figure 5 Comparison of bound intergenic regions derived from PBM data as compared with those derived from ChIP-chip[5,6]. Venn diagrams depicting the results of the comparison for Rap1 (**a**), Abf1 (**b**) and Mig1 (**c**). The Venn diagrams depict data for only those intergenic regions for which data were available for both ChIP-chip and PBMs. The asterisk indicates Rap1, Abf1 and Mig1 ChIP-chip data from Lee *et al.*[6]; the pound sign (#) indicates Rap1 ChIP-chip data from Lieb *et al.*[5]

the nutrient status of a cell with its translational capacity. Yil001w shows strong homology (BLAST E value = 4.0 × 10^{-26}) to human elongation factor 1A binding protein, implicating it in protein synthesis. *YKL082C* is thought to encode a nucleolar protein that is required for normal pre-rRNA processing and is involved in the establishment of cell polarity[23]. Expression of *YKL082C* clusters with that of several Rap1 targets identified by PBM and ChIP-chip, including *RPS27A* (ribosomal protein), *UBP10* (telomeric silencing) and *BUD22* and *BUD27* (bud site selection)[24]. Bud27 is also involved in gene expression controlled by the TOR kinase, which is known for its role in transducing the availability of nutrients into growth and ribosome synthesis.

The significantly enriched categories for the target genes derived from the Abf1 PBM data are also consistent with the known regulatory functions of Abf1 (ref. 13), including the Gene Ontology biological process categories for cell growth and/or maintenance, cell organization and biogenesis, and essentiality. Among the categories of Abf1 candidate target genes identified in this study that were not previously identified as targets by ChIP-chip, there was an enrichment for the MIPS subcellular localization functional category of the mitochondrial outer membrane, the MIPS protein complex functional category for the mitochondrial translocase complex, and the Gene Ontology biological process functional categories of nucleic acid metabolism and protein metabolism. In all, we identified 25 uncharacterized putative target genes for Abf1, approximately half of which are downstream of Abf1 sites conserved across all five *sensu stricto* species. Of note, Yhr020w shows homology to a *Drosophila melanogaster* glutamyl-prolyl-tRNA synthetase (BLAST E value = 10^{-172}). *YHR020W* is coexpressed with several other putative Abf1 targets involved in protein and nucleic acid metabolism[24].

A much more complete picture of the regulatory functions of Mig1 was possible from analysis of the PBM target genes than could be derived from the available ChIP-chip data[6]. Several known Mig1 target genes, including *DOG2* (ref. 25), *EMI2* (ref. 15), *FBP1* (ref. 26), *GAL4* (ref. 27), *GUT1* (ref. 28), *HXK1* (ref. 15), *HXT1* (ref. 15), *HXT2* (ref. 29), *JEN1* (ref. 30), *REG2* (ref. 15), *YFL054C*[15] and *YKR075C*[15], were identified only by PBMs. Among the enriched functional categories were those for C-compound and carbohydrate metabolism, carbohydrate transporters, and alcohol metabolism, all of which are consistent with the known regulatory function of Mig1 as a transcriptional repressor of genes whose products are dispensable at high levels of glucose[15]. We identified many new putative target genes for Mig1, 29 of which were previously uncharacterized, including the ORFs *YNR071C*, *YIL024C*, *YLR089C*, *YOR356W* and *YLR072W*.

Ynr071c shows strong homology (BLAST E value = 9.0 × 10^{-87}) to Gal10, which has a key role in galactose metabolism. Yil024c shows homology, albeit low (WU-BLAST2 E value = 0.09), to Sip2, a member of a family of proteins that interact with Snf1 and Snf4 and are involved in the response to glucose starvation[31]. *YLR089C* and *YOR356W* both encode proteins that are localized to the mitochondria[32] and are probably important in the catabolism of fuel molecules. Ylr089c shows homology to Bna3, which is involved in NAD biosynthesis, and to alanine aminotransferases in species ranging from plants to human (BLAST E value = 10^{-116}). These transaminases mediate the conversion of major metabolites involved in gluconeogenesis and amino acid metabolism. Yor356w shows strong homology (BLAST E value = 10^{-159}) to a human electron transfer flavoprotein-ubiquinone oxidoreductase. Notably, *YIL024C*, *YLR089C* and *YOR356W* are immediately downstream of Mig1 binding sites that are conserved in all five *sensu stricto* species. Our results also indicate that Mig1 may have a role in cholesterol biosynthesis. Mig1 shows strong homology to the human transcription factor WT1, which was recently implicated in repression of the mevalonate pathway central in cholesterol biosynthesis[33]. Similarly, Ylr072w shows homology to Atg26, a sterol 3-beta glucosyl transferase involved in sterol metabolism.

Finally, we investigated whether the collective group of target genes for each transcription factor showed concerted expression in

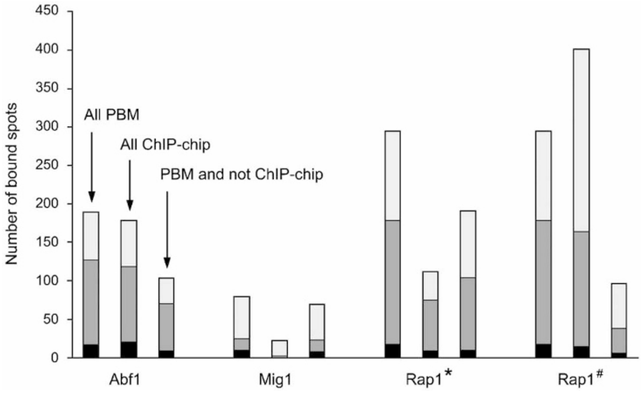


Figure 6 Cross-species sequence conservation of binding sites identified from PBM data as compared with those identified from ChIP-chip data. From left to right for a single transcription factor, bars represent all spots bound in PBMs, all spots bound in ChIP-chip and spots bound in PBMs and not ChIP-chip. The subset of bound spots with *S. cerevisiae* binding sites conserved to within two standard deviations of the motif average across all five *sensu stricto* species is shown in dark gray. The subset of *S. cerevisiae* bound spots with conserved sites 100% identical across all five species is shown in black. The remaining bound spots are shown in light gray. The asterisk indicates Rap1, Abf1 and Mig1 ChIP-chip data from Lee *et al.*[6]; the pound sign (#) indicates Rap1 ChIP-chip data from Lieb *et al.*[5]

particular experimental conditions. We hypothesized that by combining PBM and expression data, we could infer optimal conditions for transcription factor activity. We used 643 publicly available *S. cerevisiae* gene expression data sets to identify conditions in which substantial fractions of Abf1, Rap1 and Mig1 PBM target genes were differentially expressed. The conditions that resulted in the largest numbers of differentially regulated candidate target genes corresponded well with the known functions of each transcription factor (**Supplementary Note** online). For example, many Mig1 PBM target genes were downregulated by a factor of at least 2.5 in glucose and fructose, compared with other carbon sources. These results show that, together with expression profiling, PBM analysis can provide insight into the functions of particular transcription factors and identify conditions in which they are active *in vivo*. This information can also be used to prioritize conditions for ChIP-chip experiments, which require that the transcription factor under study be expressed and active.

DISCUSSION

This PBM technology allows rapid, high-throughput characterization of the DNA binding–site sequence specificities of transcription factors in a single day and can associate transcription factors with the genes they regulate. In addition to identifying enriched functional categories of known and newly identified target genes, we also identified many uncharacterized ORFs as candidate target genes of Rap1, Abf1 and Mig1. As observed for Mig1, PBM experiments will be particularly useful when ChIP-chip does not result in enough enrichment of bound fragments in the immunoprecipitated sample to permit identification of the DNA sites bound *in vivo*. ChIP-chip experiments require that the cells be in culture conditions in which the transcription factor of interest is expressed and nuclear. Furthermore, it is possible that the antibody used in ChIP-chip may not be able to detect certain classes of transcription-factor DNA binding *in vivo*, such as if the primary epitopes become inaccessible due to the formation of particular complexes at certain sites. Moreover, integrating an epitope tag on the genomic copy of the transcription factor, which allowed the use of a single antibody in the 106 ChIP-chip experiments done by Lee *et al.*[6], is not as trivial in many other organisms as it is in yeast; instead, protein-specific antibodies that are both specific and successful in chromatin immunoprecipitation are required, and the generation of such antibodies is not a trivial undertaking.

Even though the DNA in PBM experiments is not in the same state as it might be if it were to be bound by the transcription factor *in vivo*, results from PBM experiments can provide valuable data on the sequence specificity of transcription factors, particularly those that have been poorly understood or uncharacterized thus far. Carrying out ChIP-chip experiments on yeast grown under a variety of different culture conditions will help to confirm our predictions that particular sets of newly identified binding sites are indeed bound *in vivo*[34]. Furthermore, the combination of PBM data with mRNA expression data, ChIP-chip data, protein-protein interaction data and existing genetic and biochemical data in the literature will contribute to more detailed models of gene regulatory networks in yeast[35].

Results from PBM and ChIP-chip experiments might not correspond so closely for all proteins. Such differences may help to identify whether there are substantial *in vivo* effects due to chromatin structure or cofactors important in allowing or preventing sequence-specific binding. To look for evidence for such coregulatory mechanisms, we searched the sets of intergenic regions bound only *in vitro* or only *in vivo* for secondary DNA sequence motifs for each transcription factor. We did not find any statistically significant secondary motifs, potentially because of the many different modes by which binding of transcription factors to DNA is regulated *in vivo*. It is possible, however, that such secondary motifs exist for transcription factors not studied here.

The data presented here indicate that the PBM approach works for transcription factors with DNA-binding domains of a number of different structural classes. PBMs could also be used to study DNA-binding proteins important in other biological processes, such as DNA replication, DNA repair, genome rearrangement or modification of DNA. Because PBM experiments are highly scalable, they could be adapted for the analysis of all possible DNA sequence variants. Similarly, there are hundreds of predicted DNA binding proteins in yeast and thousands of predicted transcription factors in other genomes that could be screened for sequence-specific binding by PBM experiments. Because dozens of PBM experiments could be done in parallel in a single day, this technology provides considerable cost and time savings over other methods, which can take months to measure the effects of mutations for a large set of variant DNA-protein interactions.

The effects of different concentrations of transcription factors, protein cofactors, protein modifications, small molecule cofactors such as metabolites, or various binding conditions could be measured with PBMs. *In vitro* binding specifically by heterodimeric transcription factors can be detected with a PBM approach[36]. Similarly, PBMs could be used to distinguish the relative binding preferences of various whole or partially fractionated cell lysates, such as from various cell types, sampled at different time points or grown under different conditions.

Bioinformatic analysis of PBMs will provide more informative data than mononucleotide position weight matrices, as nucleotides of transcription-factor binding sites frequently do not act independently in binding by transcription factors[37–39]. Moreover, the vast data sets that would be generated on DNA-protein interactions by PBMs could yield the necessary data required to determine what predictive rules may exist that describe DNA recognition by sequence-specific transcription factors[40].

Finally, only a small handful of sequence-specific transcription factors have been characterized well enough to know many of the sequences that the transcription factors can and, just as importantly, cannot bind. More complete transcription-factor binding-site data will ultimately permit more accurate prediction of functional *cis* regulatory elements in the vast stretches of noncoding sequence in the genomes of both model organisms and the human genome than has been possible thus far[41].

METHODS

Synthesis of DNA microarrays. We synthesized microarrays spotted with double-stranded DNAs containing either positive or negative control binding sites for Rpn4 for the PBM proof-of-principle experiments with CBP-FLAG-Rpn4 essentially as described previously[11]. Exact methods and oligonucleotides are described in **Supplementary Methods** online. We synthesized whole-genome yeast intergenic microarrays essentially as described previously[3].

Expression and purification of yeast transcription factors. We created N-terminal CBP-FLAG fusions of *RPN4* by cloning *RPN4* into the pCAL-n-FLAG vector (Stratagene). We verified the full-length sequences of the resulting CBP-FLAG-RPN4 fusion constructs to ensure that no mutations had been introduced during cloning. We transformed BL21-Gold(DE3)pLysS *E. coli* (Stratagene) with the verified CBP-FLAG-RPN4 constructs and expressed them by inoculating Luria-Bertani medium containing 50 μM zinc acetate and 50 μg ml^{-1} carbenicillin with an overnight culture (1:20 dilution), growing at 30 °C to an A_{600} of 0.3–0.5, and then inducing it with 1 mM isopropyl-β-D-thiogalactopyranoside to an A_{600} of 1.0. We stored cell pellets at –80 °C and then thawed them on ice and lysed them with CelLyticB Bacterial Cell Lysis Extraction Reagent (Sigma) containing 50 μM zinc acetate. We purified the CBP-FLAG fusion proteins with anti-FLAG M2 affinity gel (Sigma) and then quantified them. We verified sequence-specific binding of the purified Rpn4 fusion protein with EMSAs using probes containing the consensus PACE site[17] (data not shown). Purified proteins were stored at –80 °C until use.

We produced N-terminal GST-His$_6$ fusions of Rap1, Abf1 and Mig1 essentially as described previously[42]. We expressed the fusion proteins in *S. cerevisiae*, purified them individually with glutathione beads (Amersham), concentrated using Microcon YM-30 filters (Millipore), and then quantified them. Purified proteins were stored at −80 °C until use.

PBM experiments. We carried out PBM experiments and SybrGreen I staining of the DNA microarrays in triplicate, essentially as described previously[11]. We thawed previously purified proteins on ice and diluted them to a final concentration of 20 nM in a 100-μl protein-binding reaction mixture consisting of phosphate-buffered saline (PBS), 50 μM zinc acetate, 2% (w/v) nonfat dried milk, 51.3 ng μl^{-1} salmon testes DNA (Sigma) and 0.2 μg μl^{-1} bovine serum albumin. We preincubated this protein-binding reaction for 1 h at room temperature. We pre-wet microarrays in PBS and 0.01% Triton X-100 and then blocked them with 2% milk in PBS for 1 h. We washed the blocked microarrays once with PBS and 0.1% Tween 20, and then once with PBS, 50 μM zinc acetate and 0.01% Triton X-100. We then applied the preincubated protein-binding mixtures to the microarrays and allowed binding to proceed for 1 h. We then washed the microarrays once with PBS, 50 μM zinc acetate and 0.5% Tween 20, and then once with PBS, 50 μM zinc acetate and 0.01% Triton X-100. We diluted Alexa 488–conjugated rabbit polyclonal antibody to GST (Molecular Probes) or Cy3-conjugated mouse M2 monoclonal antibody to FLAG (Sigma) in PBS and 50 μM zinc acetate containing 2% milk, preincubated them for at least 30 min and applied to the microarray. After incubation for 1 h, we washed the microarrays three times with PBS, 50 μM zinc acetate and 0.05% Tween 20 and once with PBS and 50 μM zinc acetate. The slides were then spun dry and stored in a closed box until being scanned.

Microarray imaging and data analysis. All whole-genome yeast intergenic microarrays were from the same print run, so as to minimize variation. We typically scanned (GSI Lumonics ScanArray 4000 or ScanArray 5000) the labeled PBMs and the SybrGreen I–stained microarrays at three to six different laser power intensities or photomultiplier tube gain settings per microarray; this allowed us to capture signal intensities for even very low signal intensity spots and ensured that we captured subsaturation signal intensities for each of the spots on the microarray[11]. We scanned microarrays using appropriate lasers and filter sets, essentially as described previously[11].

We quantified microarray TIF images using GenePix Pro version 3.0 software (Axon Instruments). We calculated background-subtracted median intensities using the median local background. We used masliner (MicroArray Spot LINEar Regression) software to calculate the relative signal intensities over the full series of laser power (or photomultiplier tube gain) setting scans in a semiautomated fashion. masliner combines the linear ranges of multiple scans from different scanner sensitivity settings onto an extended linear scale[11,43]. This resulted in final PBMs and SybrGreen I–stained microarrays having fluorescence intensities that spanned five to six orders of magnitude.

We filtered the resulting microarray data with a number of quality control criteria so that only data from high-quality spots were retained. First, for each of the triplicate microarrays, we removed data corresponding to any flagged spots (*i.e.*, spots that had dust flecks, etc.). We normalized data from each of the triplicate microarrays according to total signal intensity, so that the average spot intensity was the same for all three slides. Then we separated the data in each individual slide into sectors, according to their local region on the slide; for the whole-genome yeast intergenic arrays, we separated the spots into the 32 subgrids of the printed microarray. We then normalized the data again so that the mean spot intensity was the same over all the sectors; this served to normalize for any region-specific inhomogeneities in the background and also binding and labeling reactions. Any spots with s.d./median values >2 (*i.e.*, spots with highly variable pixel signal intensities) were filtered out. We averaged the background-subtracted, normalized signal intensities for all spots with reliable data in at least two of the three replicate microarray, calculated the means and standard deviations and filtered out any spots with a coefficient of variation (s.d./mean) >1 over the replicate microarrays. We treated the SybrGreen I microarray data exactly the same way, except that we also filtered out any spots that had $<50\%$ pixels with signal intensities >2 s.d. beyond the median background signal intensity, as these spots presumably did not have enough DNA present to allow accurate quantification of signal intensities. For the Rap1, Abf1 and Mig1 PBM datasets, ~91–96% of 6,723 unique spots passed these criteria. A detailed description of these quality control filters is available in **Supplementary Methods** online.

We carried out subsequent analyses with Perl scripts written by M.F.B. We calculated the fractional signal intensity of each spot relative to the total signal intensity on the microarray. We then calculated the log$_2$ ratio of the mean PBM signal intensity divided by the mean SybrGreen I signal intensity and created a scatter plot of the log ratio versus the spots' SybrGreen I signal intensities. Although we expect that the log ratio should be independent of DNA concentration, we found that higher DNA concentrations, as determined by higher SybrGreen I signal intensities, seem to bind proportionately less protein. To restore the independence of log ratio and SybrGreen I intensity, we fit the scatter plot with a locally weighted least squares regression using the LOWESS function[44] of the R statistics package (smoothing parameter = 0.5). We subtracted the value of the regression at each spot from its log ratio, yielding a modified log ratio that is independent of DNA concentration. We then plotted the distribution of all log ratios as a histogram (bin size = 0.05), which for the distributions of Rap1, Abf1 and Mig1 resembled a Gaussian distribution with a heavy tail. We determined the mode of the distribution by searching for the window of nine bins with the highest number of spots and taking the middle bin. We then reflected all values less than the mode and fit these values to a Gaussian function using the Mathematica software package (Wolfram Research). This gave the mean and standard deviation of the distribution, and the mean was used to adjust the log ratios so that the peak was centered on zero. We calculated a *P* value for each individual spot based on the magnitude of its log ratio relative to the standard deviation of the Gaussian distribution, using the normal error integral. To correct for multiple hypothesis testing, we adjusted all individual *P* values to a modified significance level using the Modified Bonferroni Method[38,45]. For significance testing of the PBM data, we used an initial $\alpha = 0.001$, which corresponded to α' equal to $\sim 1.5 \times 10^{-7}$ for the highest-ranking test case, as we were typically evaluating ~6,400 unique spots.

DNA motif finding and group specificity score. We used BioProspector[16] to analyze sequences for over-represented DNA sequence motifs. We chose BioProspector over other available motif-finding programs because it accepted the largest number of input sequences in construction of the transcription-factor binding-site motifs. To search for motifs that were over-represented in PBM experiments, we used all sequences from spots that had a Bonferroni-corrected *P* value ≤ 0.001 as input. To search for motifs that were over-represented in the intergenic regions bound in ChIP-chip experiments, we input either all sequences with a *P* value ≤ 0.001 (ref. 6) or all sequences with a median percentile rank ≥ 0.92 in the six replicate experiments and <0.92 in controls[5]. For each set of input sequences, we carried out separate searches at each width between 6 and 18 nucleotides to identify the highest-scoring motifs at each width. We chose the single motif with the highest group specificity score[17] to be the most significant, using the set of all sequences spotted on the microarray as the background. The group specificity score indicates the degree to which the property containing the sequence motif is specific to the input set of intergenic regions, as determined from the most significantly bound spots on the microarrays, with a smaller group specificity score indicating that the motif is more specific to the input set of spots (*i.e.*, the spots beyond a *P*-value threshold of 0.001 in the PBM data or the ChIP-chip data from Lee *et al.*[6]; the spots at or beyond the 92nd percentile rank in the ChIP-chip data from Lieb *et al.*[5]; or the randomly selected spots in the computational random controls). To assess the statistical significance of the DNA sequence motifs resulting from analysis of the PBM experiments, we carried out a set of computational negative control experiments in which we carried out identical motif searches on ten individual sets of randomly selected spots from the yeast intergenic microarrays for each transcription factor, with each random set containing the same number of sequences as the original input sets for each of the Rap1, Abf1 and Mig1 PBM data sets. The range of group specificity scores for the Rap1 control sets was 2.2×10^{-5} to 3.5×10^{-11}, with a geometric mean equal to 8.4×10^{-8}; the range of group specificity scores for the Abf1 control sets was 5.6×10^{-3} to 1.3×10^{-6}, with a geometric mean equal to 3.7×10^{-5}; and the range of group specificity scores for the Mig1 control sets was 4.8×10^{-3} to 1.8×10^{-5}, with a geometric mean equal to 4.8×10^{-4}. Thus, the Rap1, Abf1 and Mig1 motifs identified from the intergenic regions identified as bound in PBM experiments had highly significant group specificity scores as compared with the random controls. We also determined the Pearson correlation coefficients of the motifs using CompareACE[17].

To identify motifs potentially responsible for inhibition of binding or for recruitment, we carried out motif finding using BioProspector[16] and MDscan[46] on intergenic regions enriched only on PBMs or in ChIP-chip experiments, respectively, for each transcription factor. To score such potential secondary motifs, we computed group specificity scores to select those motifs enriched in intergenic regions bound only on PBMs or only in ChIP-chip relative to those bound in both types of assays. We assessed the significance of a group specificity score by comparison with scores returned from motif finding on random data sets, where the number of input sequences was the same. For identifying potential motifs under the recruitment model, we searched spots bound at P values >0.05 in PBMs and <0.05 in ChIP-chip. We selected intergenic regions for searches for motifs that might allow inhibition of transcription factor binding, if the intergenic regions had Bonferroni-corrected P values <0.001 in PBMs and >0.05 in ChIP-chip. We searched both the sequences spotted on the array and an additional 500 bp of flanking sequence on both sides of the spotted PCR amplicon, in case the ChIP-chip positives reflected transcription-factor binding to a site near the spotted intergenic sequence that then hybridized because of complementary flanking sequence due to the sonication protocol. After searching for candidate motifs, we used a CompareACE[17] cutoff of 0.7 both for merging similar discovered motifs and for identifying matches to previously known motifs.

EMSAs. We carried out EMSAs essentially in accordance with manufacturer's protocols for the LightShift Chemiluminescent EMSA Kit (Pierce). We synthesized complementary biotinylated DNA oligonucleotides, each 45 bp in length (Integrated DNA Technologies) such that they contained the predicted Rap1 binding site, flanked by its native sequence from the given intergenic region. We also synthesized a positive control probe containing a known Rap1 binding site and a negative control probe lacking a Rap1 binding site and used them in EMSAs. A list of the oligonucleotide sequences and detailed protocols used in constructing the EMSA probes is available in **Supplementary Methods** online.

Analysis of functional category enrichment. Analysis of a group of genes for enrichment for a particular functional annotation has previously been used to analyze sets of yeast genes that comprise particular gene expression clusters[22]. We used the web-based tool FunSpec for the statistical evaluation of the groups of genes downstream of the 'bound' intergenic regions, for groups of over-represented gene and protein categories with respect to existing functional category information from a number of public and published databases[47]. Like the group specificity score described above[17], FunSpec uses the hypergeometric distribution to calculate a P value for functional category enrichment[17,22].

Analysis of cross-species sequence conservation. We searched for conserved putative binding sites in the five sequenced genomes of the yeast *sensu stricto* clade: *S. cerevisiae*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus* and *S. paradoxus*. Our searches were limited to the aligned regions in the MultiZ multiple sequence alignment downloaded from the University of California Santa Cruz Genome Browser. Regions aligned between *S. cerevisiae* and each of the other four species were separately mapped onto the *S. cerevisiae* chromosomal coordinates. We used ScanACE[17] to search all five genomes for sequence matches within two standard deviations of the motif identified from PBM experiments. In our first approach, a site was considered conserved if its ScanACE score was within two standard deviations of the motif average[21] that we determined from the set of regions passing a P-value threshold of 0.001 in PBMs and if its relative position in each genome differed by no more than 15 bp. In our second approach, a site was considered exactly conserved if it satisfied the previous conditions and was identical in all five species at each of the informative positions. Here, for 'exact' conservation we used a very strict measure of sequence conservation, in which we required 100% sequence identity at all informative nucleotide positions of the binding site (9, 9 or 12 positions for Mig1, Abf1 or Rap1, respectively) in all five species. We defined informative positions to be those with an information content of greater than 0.5 bits in the PBM-derived motif. Analyses were done with Perl scripts written by M.F.B.

Analysis of correlation of target genes with gene expression data. We normalized gene expression data from 643 yeast expression microarray experiments across a variety of culture conditions[48] so that the $\log_2$ of the relative change in each microarray had a mean of 0 and standard deviation of 1. To identify conditions under which a particular transcription factor either activated or failed to repress transcription, we calculated the fraction of putative target ORFs whose expression increased by a factor of at least 2.5 for each individual condition. Similarly, to find conditions in which a transcription factor acted as a repressor or failed to activate transcription, we calculated the fraction of putative target ORFs whose expression decreased by a factor of at least 2.5 for each condition. We assessed significance by comparison with 100 sets of randomized ORFs, matched in size to the lists of target genes for each transcription factor. Each condition was assigned a score equal to the percentage of genes in each set that was upregulated, and separately downregulated, in the corresponding gene expression data set. We used the single highest score over all conditions for all random sets as our significance threshold. Any condition for which a larger fraction of predicted target genes was upregulated or downregulated was considered significant with $P < 0.01$. For this analysis, we considered ORFs to be candidate target genes if they were no more than 500 bp downstream of an intergenic region bound in PBMs at a P-value threshold of 0.001.

Analysis of candidate target genes' sequence homologies. We used the BLASTP search tool at the *Saccharomyces* Genome Database web server to analyze the sequence homologies of candidate target genes in *S. cerevisiae*, which resulted in WU-BLAST2 E values, and also in all organisms at the National Center for Biotechnology Information, which resulted in BLAST E values.

URLs. Additional data are available from http://the_brain.bwh.harvard.edu/publications.html.

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

ACKNOWLEDGMENTS
We thank T. Volkert and T. Lee for synthesis of the whole-genome yeast intergenic microarrays, B. Huber for assistance with motif searches, L. Campbell and M. Blayney for technical assistance and M. Chou and A. Philippakis for discussion. This work was supported by National Institutes of Health grants from the National Human Genome Research Institute to M.L.B. and to R.Y. and from the National Institute of General Medical Sciences to M.S. M.L.B. was also supported by a Pharmaceutical Research and Manufacturers of America Foundation Informatics Research Starter Grant and an HST Taplin Award. S.M. was supported in part by the MIT Class of 1973 Undergraduate Research Opportunities Program Fund and an MIT Bioengineering Undergraduate Research Award. M.F.B. was supported in part by a Graduate Research Fellowship from the National Science Foundation.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 7 June; accepted 18 October 2004
Published online at http://www.nature.com/naturegenetics/

1. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
2. Wodicka, L., Dong, H., Mittmann, M., Ho, M.H. & Lockhart, D.J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **15**, 1359–1367 (1997).
3. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
4. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
5. Lieb, J.D., Liu, X., Botstein, D. & Brown, P.O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat. Genet.* **28**, 327–334 (2001).

6. Lee, T. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
7. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
8. MacBeath, G. & Schreiber, S.L. Printing proteins as microarrays for high-throughput function determination. *Science* **289**, 1760–1763 (2000).
9. Ito, T. *et al.* Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. USA* **97**, 1143–1147 (2000).
10. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
11. 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).
12. Linnell, J. *et al.* Quantitative high-throughput analysis of transcription factor binding specificities. *Nucleic Acids Res.* **32**, e44 (2004).
13. Planta, R.J. Regulation of ribosome synthesis in yeast. *Yeast* **13**, 1505–1518 (1997).
14. Konig, P., Giraldo, R., Chapman, L. & Rhodes, D. The crystal structure of the DNA-binding domain of yeast RAP1 in complex with telomeric DNA. *Cell* **85**, 125–136 (1996).
15. Lutfiyya, L.L. *et al.* Characterization of three related glucose repressors and genes they regulate in *Saccharomyces cerevisiae*. *Genetics* **150**, 1377–1391 (1998).
16. Liu, X., Brutlag, D. & Liu, J. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* **2001**, 127–138 (2001).
17. Hughes, J.D., Estep, P.W., Tavazoie, S. & Church, G.M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296**, 1205–1214 (2000).
18. Wingender, E. *et al.* TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* **28**, 316–319 (2000).
19. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
20. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**, 71–76 (2003).
21. Robison, K., McGuire, A.M. & Church, G.M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.* **284**, 241–254 (1998).
22. Tavazoie, S., Hughes, J., Campbell, M., Cho, R. & Church, G. Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).
23. Drees, B.L. *et al.* A protein interaction map for cell polarity development. *J. Cell. Biol.* **154**, 549–571 (2001).
24. Beer, M.A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004).
25. Tsujimoto, Y., Izawa, S. & Inoue, Y. Cooperative regulation of DOG2, encoding 2-deoxyglucose-6-phosphate phosphatase, by Snf1 kinase and the high-osmolarity glycerol-mitogen-activated protein kinase cascade in stress responses of *Saccharomyces cerevisiae*. *J. Bacteriol.* **182**, 5121–5126 (2000).
26. Zaragoza, O., Vincent, O. & Gancedo, J.M. Regulatory elements in the FBP1 promoter respond differently to glucose-dependent signals in *Saccharomyces cerevisiae*. *Biochem. J.* **359**, 193–201 (2001).
27. Griggs, D.W. & Johnston, M. Regulated expression of the GAL4 activator gene in yeast provides a sensitive genetic switch for glucose repression. *Proc. Natl. Acad. Sci. USA* **88**, 8597–8601 (1991).
28. Grauslund, M., Lopes, J.M. & Ronnow, B. Expression of GUT1, which encodes glycerol kinase in *Saccharomyces cerevisiae*, is controlled by the positive regulators Adr1p, Ino2p and Ino4p and the negative regulator Opi1p in a carbon source-dependent fashion. *Nucleic Acids Res.* **27**, 4391–4398 (1999).
29. Ozcan, S. & Johnston, M. Function and regulation of yeast hexose transporters. *Microbiol. Mol. Biol. Rev.* **63**, 554–569 (1999).
30. Bojunga, N. & Entian, K.D. Cat8p the activator of gluconeogenic genes in *Saccharomyces cerevisiae*, regulates carbon source-dependent expression of NADP-dependent cytosolic isocitrate dehydrogenase (Idp2p) and lactate permease (Jen1p). *Mol. Gen. Genet.* **262**, 869–875 (1999).
31. Jiang, R. & Carlson, M. The Snf1 protein kinase and its activating subunit, Snf4, interact with distinct domains of the Sip1/Sip2/Gal83 component in the kinase complex. *Mol. Cell. Biol.* **17**, 2099–2106 (1997).
32. Palecek, S.P., Parikh, A.S., Huh, J.H. & Kron, S.J. Depression of *Saccharomyces cerevisiae* invasive growth on non-glucose carbon sources requires the Snf1 kinase. *Mol. Microbiol.* **45**, 453–469 (2002).
33. Rae, F.K. *et al.* Analysis of complementary expression profiles following WT1 induction versus repression reveals the cholesterol/fatty acid synthetic pathways as a possible major target of WT1. *Oncogene* **23**, 3067–3079 (2004).
34. Harbison, C.T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
35. Hartemink, A., Gifford, D., Jaakkola, T. & Young, R. Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput.* **2002**, 437–449 (2002).
36. Doi, N. *et al.* Novel fluorescence labeling and high-throughput assay technologies for *in vitro* analysis of protein interactions. *Genome Res.* **12**, 487–492 (2002).
37. 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).
38. Bulyk, M., Johnson, P. & Church, G. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* **30**, 1255–1261 (2002).
39. Udalova, I., Mott, R., Field, D. & Kwiatkowski, D. Quantitative prediction of NF-kappa B DNA-protein interactions. *Proc. Natl. Acad. Sci. USA* **99**, 8167–8172 (2002).
40. Desjarlais, J.R. & Berg, J.M. Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc. Natl. Acad. Sci. USA* **89**, 7345–7349 (1992).
41. Philippakis, A., He, F. & Bulyk, M. ModuleFinder: a tool for computational discovery of *cis* regulatory modules. *Pac. Symp. Biocomput.* (in the press).
42. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **26**, 2101–2105 (2001).
43. Dudley, A., Aach, J., Steffen, M. & Church, G. 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).
44. Cleveland, W. & Devlin, S. Locally weighted regression: An approach to regression analysis by local fitting. *J. Am. Stat. Assoc.* **83**, 596–610 (1988).
45. Sokal, R. & Rohlf, R. *Biometry: The Principles and Practice of Statistics in Biological Research* (W. H. Freeman and Company, New York, 1995).
46. Liu, X., Brutlag, D. & Liu, J. An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.* **20**, 835–839 (2002).
47. Robinson, M., Grigull, J., Mohammad, N. & Hughes, T. FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics* **3**, 35 (2002).
48. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
49. Schneider, T.D. & Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).