Distinguishing direct versus indirect transcription factor–DNA interactions

Raluca Gordân,[1] Alexander J. Hartemink,[1] and Martha L. Bulyk[2,3,4,5]

[1]*Department of Computer Science, Duke University, Durham, North Carolina 27708, USA;* [2]*Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA;* [3]*Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA;* [4]*Harvard-MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, Massachusetts 02115, USA*

Transcriptional regulation is largely enacted by transcription factors (TFs) binding DNA. Large numbers of TF binding motifs have been revealed by ChIP-chip experiments followed by computational DNA motif discovery. However, the success of motif discovery algorithms has been limited when applied to sequences bound in vivo (such as those identified by ChIP-chip) because the observed TF–DNA interactions are not necessarily direct: Some TFs predominantly associate with DNA indirectly through protein partners, while others exhibit both direct and indirect binding. Here, we present the first method for distinguishing between direct and indirect TF–DNA interactions, integrating in vivo TF binding data, in vivo nucleosome occupancy data, and motifs from in vitro protein binding microarray experiments. When applied to yeast ChIP-chip data, our method reveals that only 48% of the data sets can be readily explained by direct binding of the profiled TF, while 16% can be explained by indirect DNA binding. In the remaining 36%, none of the motifs used in our analysis was able to explain the ChIP-chip data, either because the data were too noisy or because the set of motifs was incomplete. As more in vitro TF DNA binding motifs become available, our method could be used to build a complete catalog of direct and indirect TF–DNA interactions. Our method is not restricted to yeast or to ChIP-chip data, but can be applied in any system for which both in vivo binding data and in vitro DNA binding motifs are available.

[Supplemental material is available online at http://www.genome.org.]

An essential problem in molecular biology is the identification of DNA binding sites of transcription factors (TFs) in genomes. Small-scale experiments, such as DNase footprinting or EMSA, for identifying TF binding sites are laborious and not cost-effective for high-throughput studies. In recent years, the DNA binding specificities of TFs (for brevity, we use the term "motif" henceforth to mean a model of a TF's DNA binding specificity) have been characterized via high-throughput experimental technologies such as chromatin immunoprecipitation with microarray hybridization (ChIP-chip) (Ren et al. 2000; Iyer et al. 2001; Lieb et al. 2001) followed by computational motif discovery. Dozens of motif discovery algorithms have been developed thus far (Tompa et al. 2005), but their success in identifying motifs accurately has been limited. TF motifs are typically short and degenerate, which makes them difficult to distinguish from genomic background. An additional complication when considering in vivo TF binding data is that many factors do not act alone, but rather form complexes with other TFs and thus may bind DNA directly or indirectly, depending on the precise factors and environmental conditions.

Depending on the architecture of the TF complex, sequences bound by a complex may appear enriched in ChIP-chip experiments for all the participating TFs, although only one of them binds DNA directly. For example, the yeast TFs Mbp1 and Swi6 are known to form the MBF complex, which plays a crucial role in the regulation of the cell cycle (Koch et al. 1993). Swi6 binds Mbp1, and Mbp1 contacts DNA directly at ACGCGT sequences (Taylor et al. 2000). Another example is the yeast TF Dig1. Dig1 does not have an identifiable DNA binding domain, and a literature search does not reveal any evidence of Dig1 binding DNA directly. It is known, however, that Dig1 binds DNA indirectly as part of TF complexes together with Ste12 and Tec1 (Chou et al. 2006). In such cases where a TF does not bind DNA directly, the motifs one would expect to find enriched in a ChIP-chip experiment will correspond to interacting factors (Mbp1; Ste12 or Tec1) rather than the factor that was profiled (Swi6; Dig1).

Considering the situations above, it is not surprising that motif discovery algorithms often exhibit low accuracy on in vivo data. Especially when a TF is part of several complexes with different factors interacting directly with DNA, the sequences enriched in a ChIP-chip experiment may be a complex mixture of sequences that contain binding sites for the profiled factor and/or various interacting proteins.

Here, we analyzed 237 ChIP-chip data sets from Harbison et al. (2004) to determine the extent of direct versus indirect binding by TFs in the yeast *Saccharomyces cerevisiae*. For each ChIP-chip experiment, our method determines which motifs best explain the in vivo binding data (i.e., which motifs are significantly enriched in the ChIP-chip data set). To accurately infer direct interactions between TFs and DNA, DNA binding motifs that reflect the direct sequence preferences of TFs are needed. For this purpose, we utilized motifs for 139 yeast TFs generated from independent, in vitro protein binding microarray (PBM) experiments (Bulyk et al. 2001; Mukherjee et al. 2004; Berger et al. 2006) reported recently by Badis et al. (2008) and Zhu et al. (2009). All our analyses were performed using these 139 published, PBM-derived motifs; henceforth, we use the term "motif" to refer to PBM-derived motifs, unless otherwise indicated. Within living cells, TFs often compete with nucleosomes for DNA occupancy, so our

[5]Corresponding author.
E-mail mlbulyk@receptor.med.harvard.edu; fax (617) 525-4705.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.094144.109. Freely available online through the *Genome Research* Open Access option.

19:2090–2100 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

approach also takes into account experimentally determined high-resolution, in vivo nucleosome positioning data (Lee et al. 2007).

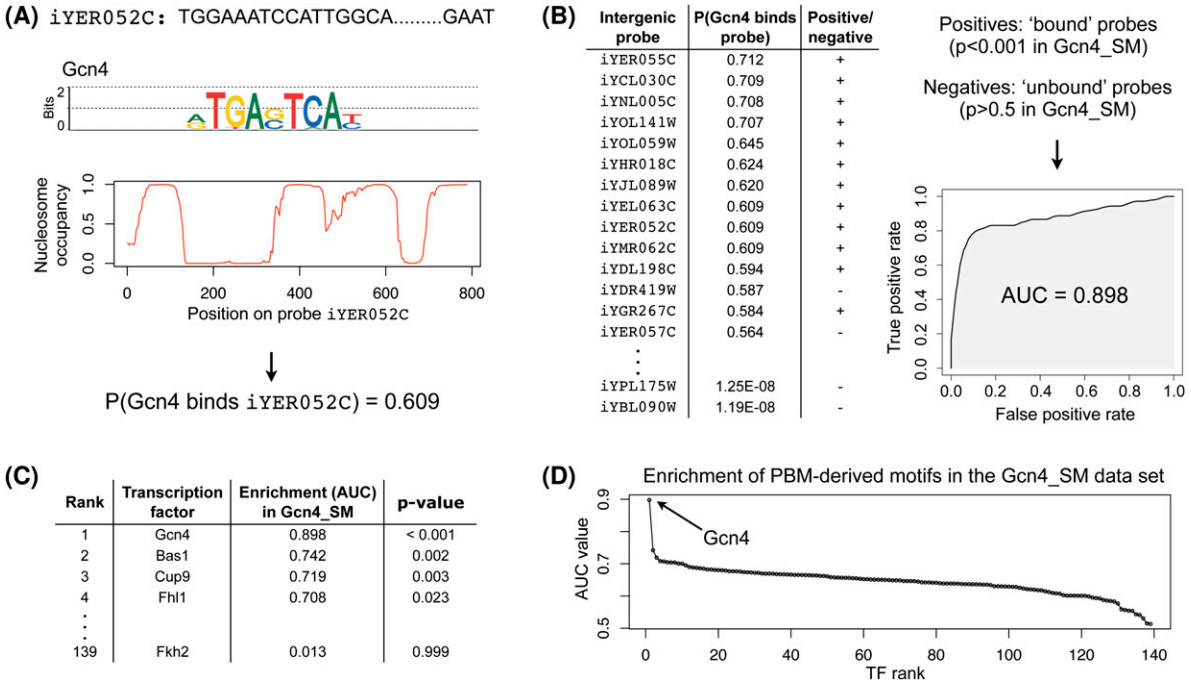
We recovered many known cases of direct and indirect DNA binding by yeast TFs. In 61 of the 128 cases in which both ChIP-chip and PBM data are available (48%), the PBM-derived motif of the factor profiled in the ChIP-chip experiment is significantly enriched in the ChIP-chip data set. In the remaining data sets, the profiled factor is not significantly enriched, suggesting that either the ChIP-chip data are too noisy or the profiled TF might associate with DNA indirectly through interaction with other proteins. Some cases in which our analysis indicates indirect TF–DNA binding are supported by experimental evidence in the literature (e.g., Dig1 binds DNA indirectly through Ste12 or Tec1), while others are novel hypotheses. Our approach is not restricted to yeast data, but could be applied to metazoan ChIP data to improve identification of direct versus indirect TF targets.

Results

Our methodology is illustrated in Figure 1. Briefly, for each of 237 ChIP-chip data sets (Harbison et al. 2004), we compute the nucleosome-aware enrichment of each of the 139 TFs for which an in vitro, PBM-derived motif was available (Badis et al. 2008; Zhu et al. 2009). We report this enrichment as the area under a receiver operating characteristic (ROC) curve (AUC), which ranges from 0 to 1, with 1 corresponding to perfect enrichment. For each of the 237 ChIP-chip data sets, we sort the 139 TFs in decreasing order of their AUC values (Fig. 1C). To assess the significance of an AUC value for a particular motif, we calculate an empirical *P*-value by generating 1000 random motifs (see Methods) and then computing their AUC values for that ChIP-chip experiment. We consider a motif's AUC value to be significant in a ChIP-chip data set if it is at least 0.65 and has an associated *P*-value ≤ 0.001.

As an example, Figure 1D shows a plot of the AUC values of all PBM-derived motifs in the ChIP-chip data set Gcn4_SM. The motif of Gcn4 (the factor profiled in that ChIP-chip experiment) is the most highly enriched, with the second ranked motif having a significantly lower AUC value. Furthermore, the only significantly enriched motif (*P*-value ≤ 0.001) is that of Gcn4. Thus, in this case we conclude that the data set Gcn4_SM can be explained by direct DNA binding of the profiled factor. Surprisingly, many ChIP-chip data sets do not exhibit this behavior; i.e., the TF profiled in the ChIP-chip experiment is not significantly enriched (see Table 2, below). A number of these cases are described in more detail below. A complete list of AUC values and associated *P*-values for all 139 PBM-derived motifs in the 237 ChIP-chip experiments is available in Supplemental Table 1. A summary of direct and indirect TF–DNA interactions, inferred from our analysis of the 237 ChIP-chip data sets, is available in Supplemental Figure 1.

The rest of this section is organized into four main parts. The first three parts discuss three categories of ChIP-chip data sets: those for which the PBM-derived motif of the profiled factor is significantly enriched, as was true for Gcn4_SM (Table 1); those for which a PBM-derived motif of the profiled factor is available, but is not significantly enriched (Table 2); and those for which a PBM-derived motif for the profiled factor is not available (Table 3). For each of these three categories, we detail a few interesting cases where independent experimental data reported in the literature support our hypothesis of indirect TF–DNA interaction. In the fourth part, we discuss the utility of incorporating in vivo nucleosome



Intergenic probe	P(Gcn4 binds probe)	Positive/ negative
iYER055C	0.712	+
iYCL030C	0.709	+
iYNL005C	0.708	+
iYOL141W	0.707	+
iYOL059W	0.645	+
iYHR018C	0.624	+
iYJL089W	0.620	+
iYEL063C	0.609	+
iYER052C	0.609	+
iYMR062C	0.609	+
iYDL198C	0.594	+
iYDR419W	0.587	-
iYGR267C	0.584	+
iYER057C	0.564	-
⋮		
iYPL175W	1.25E-08	-
iYBL090W	1.19E-08	-

Rank	Transcription factor	Enrichment (AUC) in Gcn4_SM	p-value
1	Gcn4	0.898	< 0.001
2	Bas1	0.742	0.002
3	Cup9	0.719	0.003
4	Fhl1	0.708	0.023
⋮			
139	Fkh2	0.013	0.999

Figure 1. Identification of highly enriched motifs in a ChIP-chip data set. We proceed in four steps: (*A*) For each TF with a PBM-derived motif (here, Gcn4) and each intergenic probe (here, iYER052c), we compute the probability that the TF binds that probe, as described in the Methods section. (*B*) For each TF (here, Gcn4) we rank all intergenic probes in decreasing order of the binding probability and then compute the enrichment of the motif in a ChIP-chip data set (here, Gcn4_SM) according to AUC. To calculate the AUC statistic, we defined the positive and negative sets to be the sets of intergenic regions with ChIP-chip *P*-values < 0.001 and >0.5, respectively, as calculated by Harbison et al. (2004). (*C*) For each ChIP-chip data set (here, Gcn4_SM), we ranked all TFs in decreasing order of their motif's AUC value. (*D*) We determine the significantly enriched motif(s) (here, Gcn4).

occupancy data into our analysis, as compared with the use of either in vitro nucleosome data or no nucleosome data at all.

Fewer than half of the ChIP-chip data sets are readily explained by direct DNA binding of the profiled transcription factor

We first analyzed 128 ChIP-chip data sets for which a PBM-derived motif (Badis et al. 2008; Zhu et al. 2009) is available for the profiled factor. In fewer than half of these data sets the TF profiled in the ChIP-chip experiment is significantly enriched: in 25 cases the profiled TF is the only significantly enriched factor (Table 1, left column), in 27 cases the profiled factor and factors with similar DNA binding motifs are significantly enriched (Table 1, middle column), and in nine cases the profiled factor and factors with substantially different DNA binding motifs are significantly enriched (Table 1, right column).

When the profiled TF is significantly enriched in the ChIP-chip data, we can be confident that the TF interacts directly with DNA in that condition. This is the case for ChIP-chip experiments of Abf1, Ace2, Aft2, Bas1, and 35 other TFs (Table 1). In most cases where more than one factor is significantly enriched, the enriched motifs are similar and their AUC values are almost identical. For example, in the Cbf1_YPD data set, three TFs have significant AUC values (Fig. 2A): Tye7 (AUC = 0.997), Cbf1 (AUC = 0.996), and Rtg3 (AUC = 0.991). In such cases, the enrichment of motifs for TFs other than the profiled factor may be due either to motif similarity or to an interaction between the factors. To determine whether a TF–TF interaction (here, Cbf1–Tye7, or Cbf1–Rtg3) is likely to occur, we computed the overlap between the sets of sequences bound in the ChIP-chip experiments for the TFs under consideration. If the sets of bound sequences have little or no overlap (as shown in Fig. 2C for the ChIP-chip data sets Tye7_YPD, Cbf1_YPD, and Rtg3_YPD), we conclude that the high AUC values for TFs other than the one profiled are due simply to motif similarity. This is the case for data set Cbf1_YPD: The high AUC values of Tye7 and Rtg3 are likely due to the similarity between the motifs of these two factors and the Cbf1 motif, and not to an indirect Cbf1–DNA interaction. Similar analyses for the other data sets in Table 1, middle column, showed that direct DNA binding of the profiled factor is the most likely explanation in all 27 cases.

In nine ChIP-chip experiments, the motifs of the significantly enriched TFs are not similar, although their AUC values are very close (Table 1, right column), suggesting that the enriched factors may be interacting, cooperating, or competing in the profiled conditions. Indeed, in seven of the nine cases, independent experimental evidence reported in the literature supports our

Table 1. Motifs significantly enriched in ChIP-chip data sets for which the profiled TF has a PBM-derived motif available, and this motif is significantly enriched

The profiled factor's motif is significantly enriched and		
No other motif is significantly enriched (25 data sets)	**Similar motifs are also significantly enriched (27 data sets)**	**Different motifs are also significantly enriched (nine data sets)**
Abf1_YPD: Abf1	Ace2_YPD: Swi5, Ace2	Fkh2_H2O2Hi: Hcm1, Fkh1, Mcm1, Fkh2
Aft2_H2O2Hi: Aft2	Aft2_H2O2Lo: Aft2, Aft1, Rap1	Fkh2_H2OLo: Fkh1, Mcm1, Hcm1, Fkh2
Bas1_YPD: Bas1	Bas1_SM: Bas1, Gcn4	Nrg1_H2O2Hi: Nrg1, Ecm22
Dal82_SM: Dal82	Cbf1_SM: Cbf1, Tye7, Rtg3	Sok2_BUT14: Tbs1, Phd1, Sok2
Gcn4_SM: Gcn4	Cbf1_YPD: Tye7, Cbf1, Rtg3	Ste12_BUT90: Ste12, Tec1
Hac1_YPD: Hac1	Cin5_H2O2Hi: Cin5, Yap6, Yap1	Ste12_YPD: Ste12, Mcm1
Hsf1_H2O2Hi: Hsf1	Cin5_H2O2Lo: Cin5, Yap6, Yap1	Sum1_YPD: Cup9, Ndt80, Sum1
Hsf1_H2O2Lo: Hsf1	Cin5_YPD: Cin5, Yap6, Yap1	Swi4_YPD: Swi4, Mbp1
Mbp1_H2O2Hi: Mbp1	Fkh1_YPD: Fkh2, Fkh1, Hcm1	Xbp1_H2O2Lo: Xbp1, Rds1
Mbp1_H2O2Lo: Mbp1	Fkh2_YPD: Fkh1, Fkh2, Hcm1	
Mbp1_YPD: Mbp1	Gcn4_RAPA: Gcn4, Bas1, Cup9	
Mcm1_Alpha: Mcm1	Gcn4_YPD: Gcn4, Cup9	
Mcm1_YPD: Mcm1	Gln3_RAPA: Gzf3, Dal80, Gat1, Gln3	
Pho2_SM: Pho2	Hap1_YPD: Cha4, Stb5, Oaf1, Hap1	
Reb1_H2O2Hi: Reb1	Mig1_YPD: Zms1, Mig1, Mig2, Mig3, Yml081w, Ygr067c	
Reb1_H2O2Lo: Reb1	Msn2_H2O2Hi: Ypl230w, Gis1, Rgm1, Zms1, Msn4, Rei1, Msn2	
Reb1_YPD: Reb1	Phd1_BUT90: Phd1, Sok2	
Rpn4_H2O2Lo: Rpn4	Phd1_YPD: Phd1, Sok2	
Skn7_H2O2Lo: Skn7	Pho4_Pi-: Pho4, Rtg3, Cbf1	
Stb4_YPD: Stb4	Rap1_YPD: Rap1, Aft2	
Stp4_YPD: Stp4	Rcs1_H2O2Hi: Aft1 (Rsc1), Aft2	
Ste12_Alpha: Ste12	Rcs1_H2O2Lo: Aft1 (Rsc1), Aft2, Rap1	
Tec1_BUT14: Tec1	Stb5_YPD: Hap1, Ydr520c, Stb5, Ylr278c, Oaf1, Sut2	
Tec1_YPD: Tec1	Swi5_YPD: Ace2, Swi5	
Ume6_H2O2Hi: Ume6	Tye7_YPD: Tye7, Cbf1, Rtg3	
	Ume6_YPD: Ume6, Uga3	
	Yap6_H2O2Lo: Cin5, Yap1, Yap6	
Direct DNA binding	Direct DNA binding	Direct DNA binding/coregulation[a]

In each of the three columns, the left part (e.g., Abf1_YPD) refers to a ChIP-chip data set and the right part (e.g., Abf1) refers to the TF(s) with PBM-derived motif(s) significantly enriched in that data set (i.e., with an AUC $\geq$ 0.65 and an associated *P*-value $\leq$ 0.001). Possible explanations of the ChIP-chip data are provided.

[a]We use the term ''coregulation'' to refer to any situation in which several TFs regulate, either positively or negatively, a set of genes.

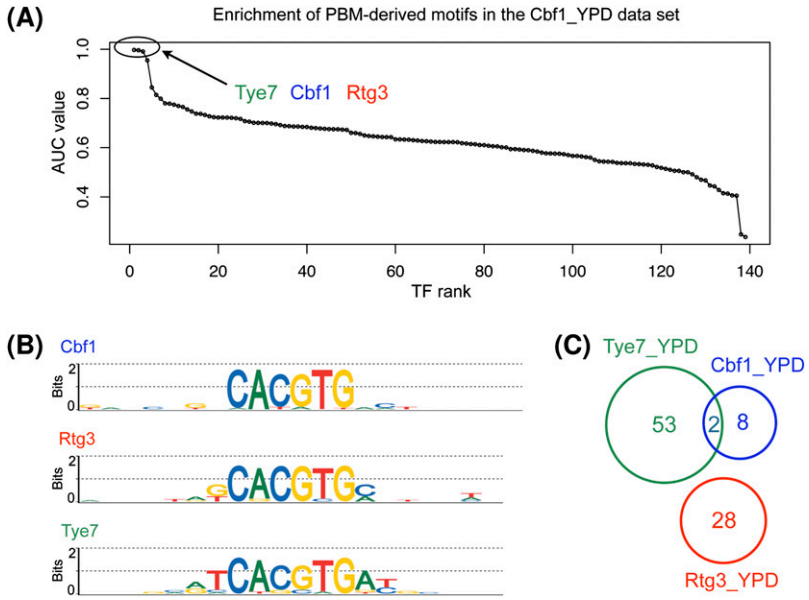


Figure 2. High-scoring motifs in the Cbf1_YPD ChIP-chip data set. (*A*) AUC values for the 139 PBM-derived motifs in the Cbf1_YPD data set. The *x*-axis shows the TF ranks, computed as in Figure 1C. (*B*) The three motifs that exhibit high AUC values in this data set: Tye7, Cbf1, and Rtg3. (*C*) Venn diagram showing the overlap among the sets of probes bound by Tye7, Cbf1, and Rtg3 in rich medium (YPD). Given the high similarity among the three motifs and the small overlap among the probes bound by the three factors, we do not consider this a case of indirect DNA binding by Cbf1.

conclusions of interaction, cooperation, or competition between significantly enriched factors and the factors profiled in the ChIP-chip experiments. The significant enrichment of Mcm1 in the ChIP-chip experiments of Fkh2 profiled in hyperoxic conditions can be explained by partial cooperation between the two factors, as described below in more detail. In the case of Sum1_YPD, Sum1 and Ndt80 have overlapping, yet distinct, sequence requirements for binding DNA, and they compete for binding to promoters containing the middle sporulation element (Pierce et al. 2003). Discussion of the other four cases supported by experimental evidence is available in the Supplemental material.

Mcm1 and Fkh2 partially cooperate in hyperoxic conditions

In the Fkh2_H2O2Hi and Fkh2_H2O2Lo data sets, we found four TFs with very high AUC values: Hcm1 (AUC = 0.894 and 0.851), Fkh1 (AUC = 0.885 and 0.874), Mcm1 (AUC = 0.880 and 0.852), and Fkh2 (AUC = 0.867 and 0.842). The motifs of Hcm1, Fhk1, and Fkh2 are very similar to each other, but different from that of Mcm1 (Fig. 3C). This suggests that the profiled factor Fkh2 and the apparently enriched Mcm1 interact or cooperate in highly and moderately hyperoxic media. Since the overlap between the probes bound by Fhk2 and Mcm1 is only partial (Fig. 3D,E), this case is probably best characterized as partial cooperation. Indeed, a literature search revealed extensive evidence for the cooperative DNA binding of Fkh2 and Mcm1 at promoters of cell-cycle genes (Hollenhorst et al. 2001).

Indirect TF–DNA interaction is suggested when the motif of the profiled TF is not significantly enriched in the ChIP-chip data

In 67 of the 128 ChIP-chip experiments for which a PBM-derived motif of the profiled factor is available, the motif is not significantly enriched in the corresponding ChIP-chip data set (Table 2). In 45 of the 67 cases, we found no motifs that explain the ChIP-chip data (Table 2, left column). At least two possible reasons could explain such cases: (1) the profiled factor binds DNA directly, but the ChIP-chip data are too noisy for this TF to appear significantly enriched, or (2) the profiled factor associates with DNA indirectly via a TF for which we did not have a PBM-derived motif available. The former might be true for data sets such as Azf1_YPD, Rds1_H2O2Hi, Sfp1_H2O2Lo, Skn7_YPD, Yap1_YPD, or Yap6_H2O2Hi, in which the profiled factor is one of the most enriched, although not enough to not pass our stringent significance criteria (AUC $\geq$ 0.65; $P \leq 0.001$).

For one additional data set—Aro80_YPD—the only significantly enriched TF is Oaf1, a factor with a DNA binding motif similar to that of the profiled factor, Aro80 (Table 2, middle column). Given the similarity between the Aro80 and Oaf1 motifs, and the fact that the sets of sequences bound in the ChIP-chip experiments of these two factors do not overlap at all, we do not consider this to be a case of indirect DNA binding by Aro80.

In the remaining 21 cases, the profiled TF does not pass the significance criteria, but factors with different DNA binding motifs do (Table 2, right column). In these cases, the ChIP-chip data might be explained by indirect association between DNA and the profiled TF, mediated by one of the factors whose motifs are significantly enriched. Supplemental Table 2 shows all the cases where our analysis indicates that a TF may bind DNA indirectly through another TF. Some interactions (Table 4, see below) are supported by independent experimental results reported in the literature, while the majority of the interactions represent novel predictions that remain to be verified in future laboratory experiments. We describe in more detail below examples for which independent experimental evidence in the literature supports the hypothesis of indirect DNA binding.

Sfp1 and Fhl1 are two factors that may bind DNA indirectly, in each case through Rap1

The PBM-derived motif of Sfp1 exhibits low enrichment in the Sfp1_SM data set, which suggests that it may not bind DNA directly, but rather as part of a TF complex. Our analysis suggests that Sfp1 binds DNA indirectly by interaction with Rap1. The Rap1 motif is the most highly enriched in the Sfp1_SM data set, with an AUC value of 0.870. The Sfp1 motif is ranked 44th, with much lower enrichment (AUC = 0.740) and an insignificant *P*-value (P = 0.597). Sfp1 is required for nutrient-dependent regulation of ribosome biogenesis (Fingerman et al. 2003) and cell size (Cipollina et al. 2008). Additionally, Sfp1 has been shown to regulate ribosomal protein (RP) gene transcription (Fingerman et al. 2003). It is not currently known whether binding of Sfp1 to RP gene promoters occurs through direct interaction with DNA or indirectly through other proteins such as Rap1 (Marion et al. 2004),

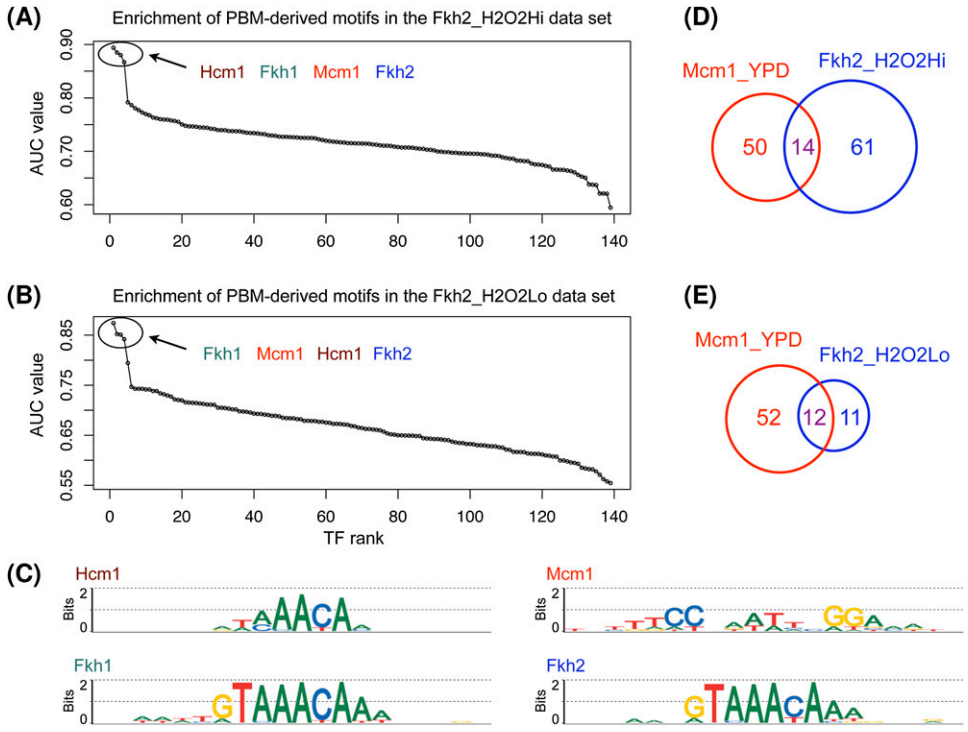


Figure 3. High-scoring motifs in the Fkh2_H2O2Hi and Fkh2_H2O2Lo ChIP-chip data sets. (*A*,*B*) AUC values for the 139 PBM-derived motifs in the two data sets. The *x*-axes show the TF ranks, computed as in Figure 1C. (*C*) Motifs significantly enriched in the two data sets. The DNA binding motif of Fkh2 was correctly identified as one of the significantly enriched motifs. In addition to Fkh2, the Hcm1, Fkh1, and Mcm1 motifs are also highly enriched. The Hcm1 and Fkh1 motifs are similar to the Fkh2 motif. Mcm1 is known to bind cooperatively with Fkh2 (Hollenhorst et al. 2001). (*D*,*E*) Venn diagrams showing the overlaps between the sets of probes bound by Fhk2 and Mcm1 in different environmental conditions.

an activator involved in many processes in *S. cerevisiae*, including transcriptional activation of RP genes (Mager and Planta 1990). Our data suggest the latter hypothesis is very likely, with Sfp1 binding RP promoters indirectly through Rap1.

Fhl1 is another factor that may bind DNA indirectly in vivo, as part of a complex with Rap1 (and also possibly Ifh1 [Schawalder et al. 2004; Wade et al. 2004]). Fhl1 was profiled by ChIP-chip after treatment with rapamycin (RAPA), in starvation medium (SM), and in rich medium (YPD) (Fig. 4). In all three data sets, the only significantly enriched motif corresponds to Rap1 (AUC = 0.819, 0.821, and 0.801; $P \leq 0.001$ in all three cases), while the Fhl1 motif ranks 10th, 12th, and 16th, with AUC values much lower than those of the Rap1 motif (AUC = 0.751, 0.758, and 0.718) and *P*-values that do not pass our significance threshold ($P = 0.077$, 0.082, and 0.114). Both Fhl1 and Rap1 associate with promoters of RP genes (Zhao et al. 2006), but Fhl1 does not appear to bind DNA directly. Rudra et al. (2005, 2007) showed that Fhl1 does not bind RP promoters directly in vitro, despite the fact that ChIP experiments clearly demonstrated that Fhl1 associates with these promoters in vivo. These investigators also found that deletion of the putative DNA binding domain of Fhl1 does not cause a significant growth defect, while mutation of a different domain (the forkhead-associated domain, which interacts with Ifh1) leads to severe defects in ribosome synthesis and growth. Additional evidence for the indirect DNA binding of Fhl1 through Rap1 comes from the work of Wade et al. (2004), who showed that although Fhl1 interacts almost exclusively with RP promoters, it does not associate with eight of the nine RP promoters that did not bind Rap1 in vivo. Furthermore, Wade et al. showed that at two of the three RP promoters tested by ChIP, the peaks of Fhl1 and Rap1 ChIP enrichment coincided. These independent experimental results support our conclusion that Fhl1 likely binds DNA indirectly in the examined culture conditions, most likely through interaction with Rap1.

Direct and indirect TF–DNA interactions can be revealed in the absence of a DNA binding motif for the profiled factor

Of the 237 ChIP-chip experiments we examined, 109 correspond to TFs for which a PBM-derived motif was not available. Although some of these factors have consensus DNA binding motifs reported in the literature, we chose not to include them in our analysis because such motifs are usually built from a small number of high-affinity DNA binding sites and may not correctly characterize medium- or low-affinity sites, which have been suggested to be abundant in vivo (Tanay 2006). Though a PBM-derived motif is not available for these factors, we can still analyze the AUC values of the 139 PBM-derived motifs to detect whether any of these motifs are significantly enriched.

In 25 of the 109 ChIP-chip data sets, we found at least one PBM-derived motif significantly enriched (Table 3). For four data sets (Table 3, left column), the significantly enriched PBM-derived motifs are similar to the DNA binding motifs of the profiled factors, as obtained from small-scale experimental studies and reported in the *Saccharomyces* Genome Database (Cherry et al. 1998); in these cases, the most likely explanation for the ChIP-chip data is direct DNA binding of the profiled factor. In the remaining 21 cases (Table 3, middle and right columns), indirect association between DNA and the profiled factor is a more likely explanation of the

Table 2. Motifs significantly enriched in ChIP-chip data sets for which the profiled TF has a PBM-derived motif available, but this motif is not significantly enriched

The profiled factor's motif is *not* significantly enriched and				
No other motif is significantly enriched (45 data sets)			**Similar motifs are significantly enriched (one data set)**	**Different motifs are significantly enriched (21 data sets)**
Aro80_SM	Mga1_YPD	Rph1_YPD	Aro80_YPD: Oaf1	Cup9_YPD: Sok2
Cha4_SM	Mig2_YPD	Rpn4_YPD		Fhl1_RAPA: Rap1
Gal4_GAL	Ndt80_YPD	Rtg3_RAPA		Fhl1_SM: Rap1
Gal4_RAFF	Nrg1_H2O2Lo	Sfp1_H2O2Lo		Fhl1_YPD: Rap1
Gal4_YPD	Oaf1_YPD	Sip4_SM		Gln3_SM: Rap1, Tbs1
Gat1_RAPA	Pdr1_H2O2Lo	Sip4_YPD		Msn4_H2O2Lo: Ecm23
Gat1_SM	Pdr1_YPD	Skn7_YPD		Msn4_Acid: Nph6a, Yox1, Smp1
Gat3_YPD	Pho2_YPD	Stp2_YPD		Nrg1_YPD: Aft2, Ypr196w, Yrm1
Gzf3_H2O2Hi	Pho4_YPD	Yap1_H2O2Lo		Pho2_H2O2Hi: Hal9, Stp4
Gzf3_RAPA	Put3_SM	Yap1_YPD		Rcs1_SM: Ypr015c, Ypr013c
Hal9_YPD	Put3_YPD	Yap6_H2O2Hi		Rtg3_H2O2Hi: Rsc30, Rds1
Leu3_SM	Rcs1_YPD	Yer130c_YPD		Rtg3_SM: Cup9
Leu3_YPD	Rds1_H2O2Hi	Yml081w_YPD		Rtg3_YPD: Gcn4, Cin5
Met32_SM	Rph1_H2O2Hi	Yox1_YPD		Sfp1_SM: Rap1
Met32_YPD	Rph1_SM	Yrr1_YPD		Skn7_H2O2Hi: Yll054c
				Smp1_YPD: Aft2
				Srd1_YPD: Fkh2
				Ste12_BUT14: Tec1
				Tec1_Alpha: Ste12
				Yap6_YPD: Phd1
				Uga3_SM: Cin5, Smp1
Indirect DNA binding via a TF for which a PBM-derived motif was not available/noisy direct DNA binding			Direct DNA binding	Indirect DNA binding

The entries in the middle and left columns are as in Table 1. Possible explanations of the ChIP-chip data are provided.

ChIP-chip data. Indeed, in several cases we found independent experimental evidence in the literature that confirms our hypothesis of indirect DNA association of the profiled TFs in certain environmental conditions (Table 4). We discuss in detail some of these cases below. A complete list of predicted TF–TF interactions is available in Supplemental Table 2.

Ste12 and Tec1 bind DNA either directly or indirectly, depending on the environmental condition

Our approach can recapitulate situations where a TF binds DNA either directly or indirectly, depending on the in vivo conditions. This is the case for Ste12 and Tec1, TFs involved in two distinct developmental programs: mating and filamentation (Chou et al. 2006). Chou and colleagues have shown that during mating—a process induced by treatment with alpha pheromone—promoters of mating genes are bound mostly by Ste12–Dig1–Dig2, but also by the Ste12–Tec1–Dig1 complex, with Ste12 binding DNA directly. During filamentation—a program induced by butanol treatment—promoters of most filamentation genes are bound by the Tec1–Ste12–Dig1 complex, with Tec1 binding DNA directly (Chou et al. 2006).

We analyzed the ChIP-chip data sets of Ste12, Tec1, and Dig1 in three environmental conditions: BUT14 (treatment with butanol for 14 h), YPD (rich medium), and Alpha (treatment with alpha pheromone). As shown in Figure 5, our results are consistent with

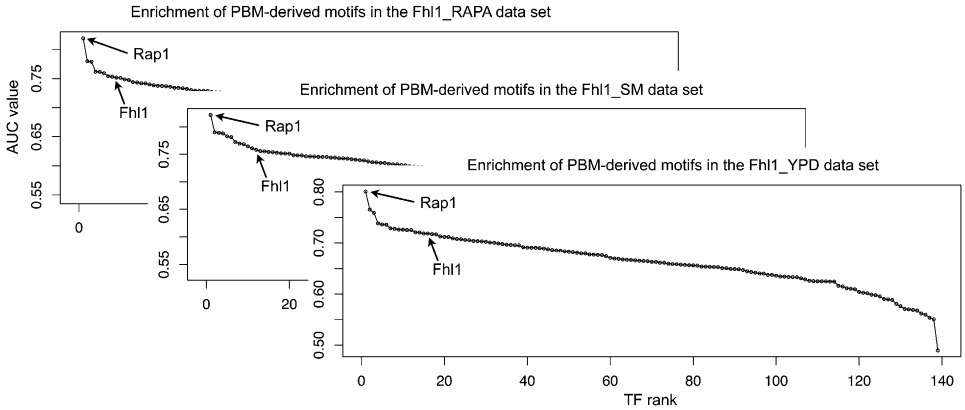


Figure 4. An example of indirect DNA association by a TF. The Rap1 motif is the only significantly enriched motif in all three Fhl1 ChIP-chip data sets: Fhl1_RAPA, Fhl1_SM, and Fhl1_YPD. The Fhl1 motif has only moderate AUC values and associated *P*-values that do not pass our significance criteria. We infer that in such cases many sequences identified as "bound" in the ChIP-chip experiments are actually indirectly bound by the profiled factor (here, Fhl1) through an interacting factor (here, Rap1).

Table 3. Motifs significantly enriched in ChIP-chip data sets for which the TF profiled by ChIP does not have an available PBM-derived motif

The significantly enriched motifs are similar to the literature motif[a] of the profiled factor (four data sets)	The significantly enriched motifs are different from the literature motif of the profiled factor (eight data sets)	A literature motif is not available for the profiled factor (13 data sets)
Ino4_YPD: Cbf1, Rtg3	Ash1_BUT14: Rds2	Dig1_Alpha: Ste12
Ino2_YPD: Cbf1, Rtg3	Dal81_RAPA: Gzf3, Gat1, Dal80, Gln3	Dig1_BUT14: Tec1
Rlm1_YPD: Smp1	Hap3_YPD: Yox1	Dig1_BUT90: Tec1
Sko1_YPD: Cst6	Hap5_SM: Gal4	Gcr2_SM: Tye7, Yap6, Cin5, Yap1, Rtg3, Cbf1
	Hap5_YPD: Nhp6a	Ixr1_YPD: Tbf1
	Mac1_H2O2Hi: Aft1, Dal82	Rlr1_YPD: Yap1
	Mot3_SM: Aft2	Ndd1_YPD: Mcm1
	Sut1_YPD: Yjl103c, Ecm22, Ylr278c, Sut2	Snt2_YPD: Stp3
		Stb1_YPD: Mbp1, Swi4
		Swi6_YPD: Mbp1, Swi4
		Ume1_H2O2Hi: Ypr013c
		Ydr026c_YPD: Reb1
		Yjl206c_H2O2Hi: Pbf1, Pbf2
Direct DNA binding	Indirect DNA binding/coregulation[b]	Indirect binding/coregulation/discovery of DNA binding motif

The entries in all three columns are as in Table 1. Possible explanations of the ChIP-chip data are provided.
[a]We use the term "literature motif" to refer to a TF's DNA binding motif as obtained from small-scale experiments and reported in the *Saccharomyces* Genome Database (Cherry et al. 1998).
[b]We use the term "coregulation" to refer to any situation in which several TFs regulate, either positively or negatively, a set of genes.

current knowledge about complexes involved in regulation of mating and filamentation: Ste12 is the only significantly enriched factor in all three experiments performed in the Alpha condition, and Tec1 is the only significantly enriched factor in all three experiments performed in the BUT14 condition. In YPD, the Ste12 and Tec1 motifs are each enriched in their respective data sets. Dig1 is not currently known to bind DNA directly, but only through Ste12 or Tec1 during mating or filamentation, respectively; thus, it is not surprising that no motif was significantly enriched in the Dig1_YPD data set.

Our method performs best when using in vivo nucleosome occupancy data

The results described thus far were obtained by integrating in vivo nucleosome occupancy data with in vivo and in vitro TF binding data. When nucleosome occupancy data are not available, one might simply consider all DNA sites to be accessible for TF binding. We performed such an analysis on the yeast ChIP-chip data sets and found that using nucleosome occupancy information significantly improves the results of our analysis. More precisely, in 60% of the ChIP-chip data sets in which a significantly enriched motif was found (Supplemental Table 4), the maximum AUC value is higher when nucleosome occupancy information is used than when it is not used. For example, the AUC value for the Rap1 motif in the Rap1_YPD data set is 0.929 when using nucleosome data, and 0.895 when nucleosome occupancy data are not used. In contrast, in 71% of the data sets in which no motif was found to be significantly enriched (Supplemental Table 5), the maximum AUC value decreased when nucleosome occupancy data were used, which suggests that any observed motif enrichment may have been due to motif matches that are nonfunctional.

We also tested our method using in vitro nucleosome sequence preference data (Kaplan et al. 2009). As expected, the overall results were slightly better than when not using any nucleosome data at all, but worse than when using in vivo data. Furthermore, for a number of TFs the results were worse when using in vitro nucleosome data than no nucleosome data at all. For example, in the cases of Abf1, Rap1, and Reb1, factors that have been shown to remodel chromatin around their binding sites (Angermayr et al. 2003; Yarragudi et al. 2004; Kaplan et al. 2009), the AUC values are lower when using in vitro data (Abf1 AUC: 0.935; Rap1 AUC: 0.865; Reb1 AUCs: 0.840, 0.957, 0.916) than when not using nucleosome data (Abf1 AUC: 0.967; Rap1 AUC: 0.894; Reb1 AUCs: 0.852, 0.982, 0.952, respectively). Since nucleosome depletion around the binding sites of these TFs in vivo can be attributable to their own action, and not to the general properties of the DNA sequence, it is not surprising that for these TFs we get worse results using in vitro nucleosome data.

Discussion

In this study, we present a systematic method to distinguish between direct and indirect TF–DNA interactions by integrating three different types of genomic data sets: ChIP-chip data on in vivo TF occupancy; PBM data on direct, in vitro DNA binding motifs of TFs; and in vivo, genomic nucleosome occupancy data. Some TFs appear to be associated with genomic sites in vivo primarily by direct DNA binding, while other TFs seem capable of binding genomic regions in vivo either directly or indirectly. Notably, of the 128 ChIP-chip data sets for which a PBM-derived motif was available for the profiled factor, fewer than half could be explained as being primarily due to direct DNA binding by the profiled factor. Moreover, the in vivo binding of a number of TFs appears to be attributable to indirect association with the genome via at least one potential interacting TF.

A caveat of our approach is that it assumes the DNA binding specificity of a TF in vivo will be the same as the specificity observed in a PBM experiment. We analyzed 21 TFs for which the PBM-derived motifs were not significantly enriched in the ChIP experiments but for which in vivo experimentally determined motifs were reported in the *Saccharomyces* Genome Database (Cherry et al. 1998), to determine whether the low enrichment may be due to the TFs having different specificities in vivo. As

Table 4. Predicted TF–TF interactions supported by independent experimental evidence in the literature

ChIP-chip experiment	No. of bound probes	Pair		TF2 motif available?[a]	AUC value of TF1 motif in ChIP-chip data set of TF2	Literature support for TF1–TF2 interaction	Similar motifs[b]
		TF1	TF2				
Dal81_RAPA	72	Gzf3 Gat1 Dal80 Gln3	Dal81	SGD	0.801 0.795 0.785 0.768	PMID: 10906145	(Gzf3, Gat1, Dal80, Gln3)
Dig1_Alpha	92	Ste12	Dig1	—	0.739	PMID: 9094309	
Dig1_BUT14 Dig1_BUT90	57 39	Tec1	Dig1	—	0.813 0.752	PMID: 16782869	
Fhl1_RAPA Fhl1_SM Fhl1_YPD	136 148 130	Rap1	Fhl1	PBM	0.819 0.821 0.801	PMID: 17452446	
Gcr2_SM	56	Tye7 Rtg3 Cbf1	Gcr2	—	0.741 0.719 0.718	PMID: 173149803[c]	(Tye7, Rtg3, Cbf1)
Hap5_SM	39	Gal4	Hap5	SGD	0.786	PMID: 11418596	
Ndd1_YPD	92	Mcm1	Ndd1	—	0.777	PMID: 14521842	
Sfp1_SM	36	Rap1	Sfp1	PBM	0.870	PMID: 15353587	
Stb1_YPD	22	Mbp1 Swi4	Stb1	—	0.763 0.749	PMID: 12832490	
Ste12_BUT14	122	Tec1	Ste12	PBM	0.811	PMID: 16782869	
Swi6_YPD	120	Mbp1 Swi4	Swi6	—	0.840 0.839	PMID: 8649372 PMID: 10747782	
Tec1_Alpha	51	Ste12	Tec1	PBM	0.679	PMID: 9234690	

[a]Specifies whether a DNA binding motif is available for TF2, either from SGD (*Saccharomyces* Genome Database), or from PBM experiments (Badis et al. 2008; Zhu et al. 2009).
[b]Groups of TFs with similar DNA binding motifs.
[c]Gcr2–Rtg3 genetic interaction.

shown in Supplemental Table 3, the in vivo motifs match the PBM-derived motifs, which suggests that the specificity of these TFs is similar in vivo and in vitro.

Previous to our study, Zhu et al. (2009) analyzed a number of ChIP-chip data sets to determine whether the profiled TFs bind DNA directly or indirectly. However, their methodology is very different from ours: For a given TF and a given intergenic sequence, Zhu and colleagues scored the sequence by summing PBM median signal intensities for each 8-mer, considering all the 8-mers with a PBM enrichment score above some threshold. In contrast, we score DNA sequences using a physically principled approach derived from GOMER (Granek and Clarke 2005), which takes into account the entire range of DNA binding affinities of the TF and thus avoids imposing thresholds on putative binding sites. Furthermore, our method can incorporate nucleosome occupancy data in a principled manner, for a more accurate distinction between direct and indirect in vivo TF–DNA interactions. Finally, we infer, and report in Table 4 and Supplemental Table 2, TF–TF interactions likely responsible for indirect DNA binding.

Liu et al. (2006) developed a method that uses nucleosome occupancy in addition to DNA binding motifs to improve detection of in vivo TF–DNA interactions. Nonetheless, Liu and colleagues incorporated nucleosome data by assuming an inhibitory effect of nucleosome occupancy and using a user-defined weight for this inhibitory effect (see Supplemental material). Moreover, Liu and colleauges applied their method to just one TF, Leu3, chosen specifically because it is known to bind DNA directly and does not have any known cofactors. Our method is much more general, and so it can be used for any TF, regardless of whether it binds DNA directly; furthermore, we were able to identify numerous cases of indirect DNA binding and associated TF–TF interactions.

The yeast ChIP-chip experiments of Harbison et al. (2004) were performed in rich medium (YPD) and 13 other culture conditions (see Methods). However, the nucleosome occupancy data used in our analysis were available only for yeast grown in YPD conditions. To analyze the importance of using nucleosome data in the same environmental condition as the ChIP-chip data, we considered a recent study by Shivaswamy et al. (2008), who reported nucleosome occupancy data for yeast grown in YPD before and after heat shock treatment (which corresponds to the YPD and HEAT conditions in the ChIP-chip data sets). Shivaswamy et al. (2008) showed that for some TFs, matches to their DNA binding motifs (MacIsaac et al. 2006) are more accessible in HEAT than in YPD. However, we found that in both of these conditions,

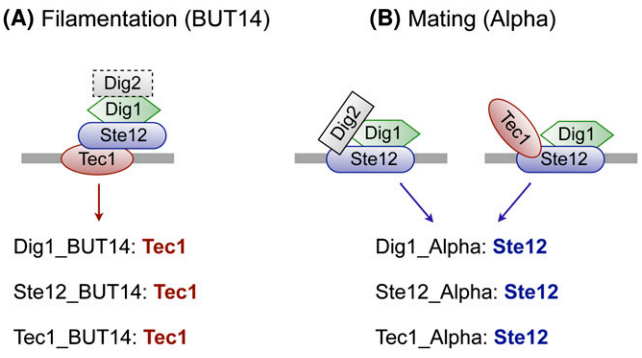


Figure 5. Direct and indirect DNA binding by Ste12 and Tec1. Ste12 and Tec1 are both involved in two developmental processes: filamentation (induced by treatment with butanol, as in the BUT14 condition) and mating (induced by treatment with the alpha pheromone, as in the Alpha condition). (*A*) During filamentation, the Tec1-Ste12–Dig1 complex binds DNA through Tec1. Our method correctly identifies Tec1 as the only significantly enriched TF in the ChIP-chip experiments where filamentation occurs. (*B*) During mating, the Ste12–Dig1–Dig2 and Ste12–Tec1–Dig 1 complexes bind DNA through Ste12. Our method correctly identifies Ste12 as the only significantly enriched TF in the ChIP-chip experiments where mating occurs.

functional DNA binding sites are in general more accessible than neighboring DNA sites (Supplemental Fig. 2), supporting our incorporation of nucleosome occupancy data in our analysis. Nevertheless, it would be preferable to use nucleosome occupancy data for yeast grown in the same environmental (and genetic) conditions as the yeast profiled by ChIP-chip. In the future, as additional high-resolution nucleosome occupancy data are generated for yeast grown in other culture conditions, such occupancy data could be easily incorporated into our analysis to provide more precise predictions of direct versus indirect binding events in the genome.

The approach described in this study is not restricted to yeast or to ChIP-chip data, but could be applied to the analysis of ChIP-seq (Johnson et al. 2007) or ChIP-PET (Wei et al. 2006) data sets for TFs in other organisms, including metazoans. With the generation of diverse PBM data sets for hundreds of metazoan TFs (Berger et al. 2008; Badis et al. 2009; Grove et al. 2009), this approach may not only distinguish direct versus indirect genomic TF binding events in vivo, but also suggest the identities of the interacting TFs.

Methods

ChIP-chip data

We used the yeast ChIP-chip data from Harbison et al. (2004), who performed 352 ChIP experiments for 207 TFs under different environmental conditions: YPD (rich medium), Acid (acidic medium), Alpha (alpha factor pheromone treatment), BUT14 (butanol treatment for 14h), BUT90 (butanol treatment for 90 min), GAL (galactose medium), H2O2Hi (highly hyperoxic), H2O2Lo (mildly hyperoxic), HEAT (elevated temperature), Pi- (phosphate deprived medium), RAFF (raffinose medium), RAPA (nutrient deprived), SM (amino acid starvation), and THI- (vitamin deprived). We use the notation *TF_cond* to refer to the ChIP-chip experiment for transcription factor "TF" under environmental condition "cond." For each ChIP-chip data set, we defined the "bound" intergenic probes to be those with a *P*-value < 0.001. We restricted our analysis to the 237 (out of 352) data sets that contained at least 10 probes bound at $P < 0.001$.

PBM-derived DNA binding motifs

Badis et al. (2008) and Zhu et al. (2009) used universal PBMs (Berger et al. 2006) to determine high-resolution in vitro DNA binding specificity data for 139 TFs. They reported PBM-derived motifs for these TFs as position weight matrices (PWMs). We used all 89 PWMs of Zhu et al. (2009) and 50 additional PWMs from Badis et al. (2008).

Nucleosome positioning data

We used in vivo nucleosome positioning information from Lee et al. (2007) to compute, for each DNA site $\boldsymbol{S}$, the probability that the site is occupied by a nucleosome. Lee et al. used micrococcal nuclease digestion followed by microarray analysis to derive a high-resolution map of nucleosome occupancy across the whole yeast *S. cerevisiae* genome. From this map we extracted, for every fourth position in the genome, the logarithm of the ratio between the signal intensity of nucleosomal DNA versus genomic DNA at that position, and then interpolated the data to obtain 1-bp resolution data. Next, we applied a logistic transformation to the log-ratio values to obtain, for each position in the genome, the probability of that position being occupied by a nucleosome (see Supplemental material for details).

Given a site $\boldsymbol{S} = S_1 \ldots S_W$ of width W and the probability of nucleosome occupancy at each position i in the site, we can compute the probability of site $\boldsymbol{S}$ being occupied by a nucleosome, or, alternatively, the probability of site $\boldsymbol{S}$ being free of nucleosomes:

$$P(S_1 \ldots S_W \text{ free}) = P(S_1 \text{ free}) \times P(S_2 \text{ free} \mid S_1 \text{ free}) \times \ldots \times P(S_W \text{ free} \mid S_{W-1} \text{ free}) \quad (1)$$

Each term $P(S_{i+1} \text{ free} \mid S_i \text{ free})$ can be written as:

$$P(S_{i+1} \text{ free} \mid S_i \text{ free}) = 1 - P(S_{i+1} \text{ occupied} \mid S_i \text{ free}) = 1 - \frac{1}{N} \times (S_{i+1} \text{ occupied}) \quad (2)$$

where N is set to 147, the average nucleosome width.

Scoring a DNA sequence according to a PWM

We scored DNA sequences using a model similar to GOMER (Granek and Clarke 2005). Other models such as MatrixREDUCE (Foat et al. 2006) or TRAP (Roider et al. 2007) could also be used to compute the probability that a TF with a particular PWM binds a DNA sequence. However, both MatrixREDUCE and TRAP use parameters that need to be trained on the ChIP-chip data. Since we want to use the model to test how well certain motifs explain the ChIP-chip data, training those motifs on the data themselves would not be appropriate.

Let $\boldsymbol{T}$ denote a TF, and ϕ denote the PWM describing the DNA binding motif of $\boldsymbol{T}$: $\phi(b,j)$ = the probability of finding base b at location j within the binding site ($b \in \{A, C, G, T\}$ and $1 \le j \le W$, where W is the width of the motif). Let ϕ_0 denote the background model, a 0th-order Markov model trained on all intergenic sequences in yeast.

Given a DNA site $\boldsymbol{S} = S_1 S_2 \ldots S_W$, we score it according to the PWM and background models, and use the ratio of the two scores as an approximation for the dissociation constant $K_d(\boldsymbol{T}, \boldsymbol{S}) = \prod_{j=1}^{W} \frac{\phi_0(S_j)}{\phi(S_j, j)}$. Next, using the fact that $K_d(\boldsymbol{T},\boldsymbol{S}) = [\boldsymbol{T}] \cdot [\boldsymbol{S}] / [\boldsymbol{T} \cdot \boldsymbol{S}]$, we can write the probability that TF $\boldsymbol{T}$ binds DNA site $\boldsymbol{S}$ as:

$$P(\boldsymbol{T} \text{ binds } \boldsymbol{S}) = \frac{[\boldsymbol{T} \cdot \boldsymbol{S}]}{[\boldsymbol{T} \cdot \boldsymbol{S}] + [\boldsymbol{S}]} = \frac{[\boldsymbol{T}]}{[\boldsymbol{T}] + K_d(\boldsymbol{T}, \boldsymbol{S})} = 1 \Big/ \left(1 + \frac{1}{[\boldsymbol{T}]} \times \prod_{j=1}^{W} \frac{\phi_0(S_j)}{\phi(S_j, j)}\right) \quad (3)$$

where the concentration of free TF, $[\boldsymbol{T}]$, is set to the dissociation constant for the site with the optimal PWM score, as in the GOMER model (Granek and Clarke 2005).

For a DNA sequence $\boldsymbol{X}$ longer than the motif width W, the probability that TF $\boldsymbol{T}$ binds $\boldsymbol{X}$ is:

$$P(\boldsymbol{T} \text{ binds } \boldsymbol{X}) = P(\boldsymbol{T} \text{ binds any } X_i \ldots X_{i+W-1}) = 1 - \prod_{i}^{n-W+1} \left(1 - 1\Big/\left(1 + \frac{1}{[T]} \times \prod_{j=i}^{i+W-1} \frac{\phi_0(X_j)}{\phi(X_j, j-i+1)}\right)\right) \quad (4)$$

Incorporating nucleosome positioning information

So far we assumed that the probability that a TF binds a DNA site depends only on the specificity of the factor for that particular site, which is a good assumption in the case of in vitro experiments. In vivo, however, many DNA regions are occupied by nucleosomes and thus are not accessible for binding by a TF. To take this into account, we first need to rewrite Equation 3 to include information about the accessibility of site $\boldsymbol{S}$:

$$\begin{aligned} P(\boldsymbol{T} \text{ binds } \boldsymbol{S}) &= P(\boldsymbol{T} \text{ binds } \boldsymbol{S} \mid \boldsymbol{S} \text{ free}) \times P(\boldsymbol{S} \text{ free}) \\ &\quad + P(\boldsymbol{T} \text{ binds } \boldsymbol{S} \mid \boldsymbol{S} \text{ occupied}) \times P(\boldsymbol{S} \text{ occupied}) \\ &= P(\boldsymbol{T} \text{ binds } \boldsymbol{S} \mid \boldsymbol{S} \text{ free}) \times P(\boldsymbol{S} \text{ free}) \end{aligned} \quad (5)$$

The second equality follows from the assumption that sites occupied by nucleosomes have zero probability of being accessed by TFs. Although a few TFs have been observed to bind nucleosomal DNA, our assumption is true for the vast majority of factors.

Taking into account nucleosome occupancy information, Equation 4 can be rewritten as Equation 6, where $P(X_i \ldots X_{i+W-1}$ free) is derived from the in vivo nucleosome occupancy data.

$$P(\boldsymbol{T} \text{ binds } \boldsymbol{X}) = 1 - \prod_{i}^{n-W+1} \left(1 - 1\Big/\left(1 + \frac{1}{[T]} \times \prod_{j=i}^{i+W-1} \frac{\phi_0(X_j)}{\phi(X_j, j-i+1)}\right) \times P(X_i \ldots X_{i+W-1} \text{ free})\right) \quad (6)$$

Given a DNA sequence, a PBM-derived motif, and the nucleosome occupancy information over that sequence, we use Equation 6 to compute the probability that the TF binds that sequence, as shown in Figure 1A for TF Gcn4 and intergenic region iYER052C.

Analyzing data from a ChIP-chip experiment

We use the probability that a TF $\boldsymbol{T}$ binds a DNA sequence $\boldsymbol{X}$ to score every intergenic probe present on the microarrays used in the ChIP-chip experiments (Harbison et al. 2004). For example, Figure 1B shows the probability of TF Gcn4 binding each yeast intergenic region. Next, for any particular ChIP-chip experiment we define two sets of intergenic probes: the positive set (i.e., the set of "bound" probes), which contains all the probes with a P-value < 0.001, and the negative set (i.e., the set of "unbound" probes), which contains all the probes with a P-value > 0.5, as calculated by Harbison et al. (2004); we did not consider probes with intermediate P-values. Using the positive and negative sets from each ChIP-chip experiment, and the probabilities that TF $\boldsymbol{T}$ binds each of the probes, we compute the enrichment of the PBM-derived motif for TF $\boldsymbol{T}$ in the ChIP-chip data by an AUC value. For each ChIP-chip experiment *TF_cond* we computed the AUC values of the 139 DNA binding motifs derived from PBM data.

Computing the statistical significance of AUC values

To assess whether the AUC value computed for a PBM-derived motif in a particular ChIP-chip data set is significant, we proceeded in three steps: (1) We randomly generated 1000 motifs by permuting the nucleotides in each column of the initial motif; (2) for each random motif, we computed its AUC value in the given ChIP-chip data set; and (3) we used the 1000 AUC values to compute an empirical P-value for the AUC of the real motif. We consider an AUC value significant if it is at least 0.65 (i.e., it explains the ChIP-chip data to some extent) and has an associated P-value ≤ 0.001 (i.e., at most one of the 1000 random motifs has an AUC value equal to or greater than the AUC value of the real motif).

Acknowledgments

We thank C. Zhu and K. Byers for sharing pre-publication yeast PBM data, and R.P. McCord for helpful discussion and critical reading of the manuscript. This work was funded by grants from NIH (R01 HG003985, R01 HG003420) to M.L.B., and by an NSF CAREER award (0347801), an Alfred P. Sloan Research Fellowship, and grants from NIH (P50 GM081883-01, R01 ES015165-01) and DARPA (HR0011-08-1-0023, HR0011-09-1-0040) to A.J.H.

References

Angermayr M, Oechsner U, Bandlow W. 2003. Reb1p-dependent DNA bending effects nucleosome positioning and constitutive transcription at the yeast profilin promoter. *J Biol Chem* **278:** 17918–17926.

Badis G, Chan ET, van Bakel H, Pena-Castillo L, Tillo D, Tsui K, Carlson CD, Gossett AJ, Hasinoff MJ, Warren CL, et al. 2008. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* **32:** 878–887.

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, et al. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science* **324:** 1720–1723.

Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW III, Bulyk ML. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **24:** 1429–1435.

Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, et al. 2008. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133:** 1266–1276.

Bulyk ML, Huang X, Choo Y, Church GM. 2001. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci* **98:** 7158–7163.

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, et al. 1998. SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res* **26:** 73–79.

Chou S, Lane S, Liu H. 2006. Regulation of mating and filamentation genes by two distinct Ste12 complexes in *Saccharomyces cerevisiae*. *Mol Cell Biol* **26:** 4794–4805.

Cipollina C, van den Brink J, Daran-Lapujade P, Pronk JT, Porro D, de Winde JH. 2008. *Saccharomyces cerevisiae* SFP1: At the crossroads of central metabolism and ribosome biogenesis. *Microbiology* **154:** 1686–1699.

Fingerman I, Nagaraj V, Norris D, Vershon AK. 2003. Sfp1 plays a key role in yeast ribosome biogenesis. *Eukaryot Cell* **2:** 1061–1068.

Foat BC, Morozov AV, Bussemaker HJ. 2006. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22:** e141–e149.

Granek JA, Clarke ND. 2005. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol* **6:** R87. doi: 10.1186/gb-2005-6-10-r87.

Grove CA, De Masi F, Barrasa MI, Newburger DE, Alkema MJ, Bulyk ML, Walhout AJ. 2009. A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* **138:** 314–327.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, MacIsaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431:** 99–104.

Hollenhorst PC, Pietz G, Fox CA. 2001. Mechanisms controlling differential promoter-occupancy by the yeast forkhead proteins Fkh1p and Fkh2p: Implications for regulating the cell cycle and differentiation. *Genes & Dev* **15:** 2445–2456.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409:** 533–538.
Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502.
Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458:** 362–366.
Koch C, Moll T, Neuberg M, Ahorn H, Nasmyth K. 1993. A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science* **261:** 1551–1557.
Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C. 2007. A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* **39:** 1235–1244.
Lieb JD, Liu X, Botstein D, Brown PO. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* **28:** 327–334.
Liu X, Lee CK, Granek JA, Clarke ND, Lieb JD. 2006. Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome Res* **16:** 1517–1528.
MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E. 2006. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* **7:** 113. doi: 10.1186/1471-2105-7-113.
Mager WH, Planta RJ. 1990. Multifunctional DNA-binding proteins mediate concerted transcription activation of yeast ribosomal protein genes. *Biochim Biophys Acta* **1050:** 351–355.
Marion RM, Regev A, Segal E, Barash Y, Koller D, Friedman N, O'Shea EK. 2004. Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *Proc Natl Acad Sci* **101:** 14315–14322.
Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML. 2004. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* **36:** 1331–1339.
Pierce M, Benjamin KR, Montano SP, Georgiadis MM, Winter E, Vershon AK. 2003. Sum1 and Ndt80 proteins compete for binding to middle sporulation element sequences that control meiotic gene expression. *Mol Cell Biol* **23:** 4814–4825.
Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306–2309.
Roider HG, Kanhere A, Manke T, Vingron M. 2007. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* **23:** 134–141.
Rudra D, Zhao Y, Warner JR. 2005. Central role of Ifh1p–Fhl1p interaction in the synthesis of yeast ribosomal proteins. *EMBO J* **24:** 533–542.
Rudra D, Mallick J, Zhao Y, Warner JR. 2007. Potential interface between ribosomal protein production and pre-rRNA processing. *Mol Cell Biol* **27:** 4815–4824.
Schawalder SB, Kabani M, Howald I, Choudhury U, Werner M, Shore D. 2004. Growth-regulated recruitment of the essential yeast ribosomal protein gene activator Ifh1. *Nature* **432:** 1058–1061.
Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, Iyer VR. 2008. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol* **6:** e65. doi: 10.1371/journal.pbio.0060065.
Tanay A. 2006. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res* **16:** 962–972.
Taylor IA, McIntosh PB, Pala P, Treiber MK, Howell S, Lane AN, Smerdon SJ. 2000. Characterization of the DNA-binding domains from the yeast cell-cycle transcription factors Mbp1 and Swi4. *Biochemistry* **39:** 3943–3954.
Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23:** 137–144.
Wade JT, Hall DB, Struhl K. 2004. The transcription factor Ifh1 is a key regulator of yeast ribosomal protein genes. *Nature* **432:** 1054–1058.
Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, Shahab A, Yong HC, Fu Y, Weng Z, et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124:** 207–219.
Yarragudi A, Miyake T, Li R, Morse RH. 2004. Comparison of ABF1 and RAP1 in chromatin opening and transactivator potentiation in the budding yeast *Saccharomyces cerevisiae*. *Mol Cell Biol* **24:** 9152–9164.
Zhao Y, McIntosh KB, Rudra D, Schawalder S, Shore D, Warner JR. 2006. Fine-structure analysis of ribosomal protein gene transcription. *Mol Cell Biol* **26:** 4853–4862.
Zhu C, Byers K, McCord R, Shi Z, Berger M, Newburger D, Saulrieta K, Smith Z, Shah M, Radhakrishnan M, et al. 2009. High-resolution DNA binding specificity analysis of yeast transcription factors. *Genome Res* **19:** 556–566.

Received March 27, 2009; accepted in revised form July 29, 2009.