Supporting online material for:

Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*

Tong Ihn Lee*, Nicola J. Rinaldi*, François Robert*, Duncan T. Odom, Ziv Bar-Joseph, Georg K. Gerber, Nancy M. Hannett, Christopher T. Harbison, Craig M. Thompson, Itamar Simon, Julia Zeitlinger, Ezra G. Jennings, Heather L. Murray, D. Benjamin Gordon, Bing Ren, John J. Wyrick, Jean-Bosco Tagne, Thomas L. Volkert, Ernest Fraenkel, David K. Gifford, Richard A. Young

*These authors contributed equally to this work

METHODS

Epitope Tagging of Strains

Regulators were tagged by inserting multiple copies of the Myc epitope coding sequence into the normal chomosomal loci of these genes. The plasmid pWZV88 (1) was used as a template to generate PCR products containing the Myc epitope coding sequence and a selectable marker (TRP) flanked by homologous regions designed to recombine at the 3' end of the targeted transcriptional regulator. PCR products were transformed into the W303 strain Z1256. Clones were selected for growth on TRP- plates. Insertion of the epitope coding sequence was confirmed by PCR and expression of the epitope-tagged protein was confirmed by Western blotting using an anti-Myc antibody.

Genome-wide Location Analysis

The genome-wide location analysis method we have developed allows protein-DNA interactions to be monitored across the entire yeast genome by combining a modified Chromatin Immunoprecipitation (ChIP) procedure, which has been previously used to study *in vivo* protein-DNA interactions at one or a small number of specific DNA sites, with DNA microarray analysis. Briefly, cells containing a copy of an epitope tagged regulator were fixed with formaldehyde (1% final concentration) and then harvested by centrifugation. Cells were lysed with glass beads and the resulting cell lysate was sonicated to shear DNA. DNA fragments representing promoter regions crosslinked to a protein of interest were enriched by immunoprecipitation with an anti-epitope antibody. After reversal of the crosslinking, the enriched DNA was amplified and labeled with a fluorescent dye by ligation-mediated PCR (LM-PCR). A sample of DNA that had not been enriched by immunoprecipitation was subjected to LM-PCR in the presence of a different fluorophore, and both IP-enriched and unenriched pools of labeled-DNA were hybridized to a single DNA microarray containing all yeast intergenic sequences. Slides were then scanned. For each factor, three independently grown cell cultures were processed and scanned to generate binding information.

Data analysis

Scanned images were analyzed with either ArrayVision 5.0 or GenePix 3.0, to obtain background-subtracted intensity values. Each spot was bound by both immunoprecipitation(IP)-enriched and unenriched DNA which were labeled with different fluorophores. Consequently, each spot yielded fluorescence intensity information in two channels, corresponding to immunoprecipitated DNA and genomic DNA. To account for background hybridization to slides, the median intensity of a set of control blank spots was subtracted. To correct for different amounts of genomic and immunoprecipiated DNA hybridized to the chip, the median intensity value of the IP-enriched DNA channel was divided by the median of the genomic DNA channel, and this normalization factor was applied to each intensity in the genomic DNA channel. Next, we calculated the log of the ratio of intensity in the IP-enriched channel to intensity in the genomic DNA channel for each intergenic region across the entire set of hybridization experiments. To account for systematic biases introduced by the immunoprecipitation, all of the log ratios for a specific intergenic region were then normalized by subtracting the average log ratio for that intergenic region. Adjusted intensity values for the IP-enriched channel were calculated from these normalized ratios. A whole chip error model (2) was then used to calculate confidence values for each spot on each chip, and to combine data for the replicates of each experiment to obtain a final average ratio and confidence for each intergenic region. Further details including protocols and information about spot to gene assignment can be found at http://web.wi.mit.edu/young/regulator_network

QUALITY CONTROL

Several methods have been used to estimate the quality of the genome-wide data described here. We discuss below our estimates of false-positive and false-negative rates in data reported at a p-value threshold of 0.001.

Error Model

Error models assign a confidence measure (p-value) to the results of microarray data and are useful because biological experiments and microarray technology produce noisy data. Furthermore, because DNA binding proteins are believed to be in an equilibrium between the bound and unbound state, the p-value provided by error models is likely to provide a more relevant criterion of binding than a strict "bound or not bound" assignment. Nonetheless, it is sometimes necessary to assign a threshold to the p-value for the purposes of analyzing, discussing and generating regulatory models from the data. The influence of assigning various p-value thresholds on the number of binding interactions was described in Figure 1b.

We have generally used a p-value threshold of 0.001 to analyze, discuss and generate regulatory models from the genome-wide location data described here. We selected this value because it minimizes false positive results, as discussed in detail below. This comes at the expense of an increase in false negative results. The accuracy of genome-wide location data reported with confidence levels of 0.001 or better has been confirmed using several approaches, which are discussed in more detail below.

Experimental confirmation of data

Conventional, independent chromatin immunprecipitation experiments conducted in our laboratory at a gene-specific level have confirmed 93 of 99 binding interactions (involving 29 different regulators) that were identified by location analysis data at a threshold p-value of 0.001. This suggests that our empirical rate of false positives is 6%. A sample of regulator-gene interactions confirmed by gene-specific PCR is shown in Figure S1. A summary of the regulators-gene interactions confirmed in these experiments can be found in Table S1.

A second approach used to estimate the false positive rate was to conduct control experiments with arrays hybridized with differently labeled but otherwise identical pools of control DNA. We can estimate a false positive rate by finding the number of positives generated from sets of three of these control DNA vs. control DNA arrays. We generated 1,000 randomized sets of three arrays from the control DNA arrays. The results indicate that an average of 3.7 out of ~6500 intergenic regions qualify as significantly enriched using the 0.001 threshold. The average number of intergenic regions that qualify as significantly enriched using the 0.001 threshold in actual experiments is 38, so from this we estimate an average false positive rate of 10%.

Literature confirmation of the data

Where there is extensive literature describing previous evidence for binding of regulators to specific promoters *in vivo*, we have compared the previous reports with data from genome-wide location analysis. We find that the location data generally agrees with the published literature. For example, we previously noted that the nine cell cycle

regulators Mbp1, Swi4, Swi6, Mcm1, Fkh1, Fkh2, Ndd1, Swi5 and Ace2 form 50 regulator-gene interactions based on published genetic, biochemical or direct *in vivo* binding experiments (3). The location data predicts 41 of these at the 0.001 threshold and 46 at the 0.005 threshold. These data have been summarized in Table S2.

The literature comparison does not allow us to estimate an accurate false-positive rate because the literature is incomplete relative to the dataset we have generated. The literature does identify some regulator-gene interactions that we have not observed, indicating that some interactions are not reported by the location data at a p-value threshold of 0.001. We discuss false-negatives in more detail below.

Quantifying false negatives

The 0.001 p-value threshold was used to minimize false positive reports, with the expectation that this threshold may result in an underestimate of regulator-DNA interactions. Consequently, it is useful to estimate the number of regulator-DNA interactions that are missed by our analysis at the 0.001 threshold. The determination of a true false negative rate in our dataset is not feasible. To establish this false negative rate, we would require perfect knowledge of the interactions that actually occur. We could then calculate how faithfully the microarray experiments recapitulate the ground truth.

Despite this challenge, we used the following approach to obtain a rough estimate the number of binding interactions that are not reported in the genome-wide location data using the 0.001 p-value threshold. We used gene-specific PCR analysis with selected regulators to test the results predicted at each of the different p-value thresholds. We could then determine how frequently a regulator-gene interaction is correctly predicted at p values greater than 0.001. The percentage of confirmed interactions at these less stringent thresholds indicates how many interactions we are potentially missing.

We selected two different regulators, Nrg1 and Stb1, for further testing. For each regulator, we selected sets of genes with p-values closest to one of four thresholds (0.001, 0.005, 0.01, 0.05) and performed chromatin IP and gene-specific PCR. The data are summarized in Figure S2 and specific interactions are listed in Table S3. Using these results as a guide, we estimate that at least 2,300 additional genuine regulator-gene interactions exist among our results at all p-values above 0.001. While comparisons to published literature suggest a false negative rate of 18% (as we do not identify 9 out of 50 published interactions for cell cycle regulators as described above), we feel the estimates from this direct approach are more accurate.

GENOME COVERAGE

We have used genome-wide location analysis to identify genomic binding sites for many transcriptional regulators in living yeast cells under a single growth condition. Here we discuss the limitations of the current location dataset in terms of the extent to which this study has covered all yeast genes and the factors involved in obtaining more complete coverage.

We found that the promoter regions for 2343 out of 6270 yeast genes (37%) are bound by one or more of the 106 transcriptional regulators in yeast cells grown in rich medium when the location data is reported using a p-value threshold of 0.001. Since more than half of yeast genes were not assigned a regulator by this analysis, we considered what factors might allow us to more fully populate a map of the yeast transcriptional regulatory network in the future. The factors we considered were limitations attributable to genome-wide location analysis technology, the rates of false negatives due to the high confidence threshold, regulators that have yet to be identified, and the contribution of extracellular environment.

Genome-wide location analysis technology

We considered the possibility that the genes that are not bound by the set of 106 regulators may be underrepresented in the present genome-wide location dataset due to defects in the DNA microarray. The DNA microarrays used in this study were produced as described previously (4), and were subjected to a series of quality control tests.

We assayed the ability of individual spots on the arrays to report hybridization of labeled genomic DNA, and found that 95% of the spots produce effective signals in this assay. We found that approximately 5% of the 6600 intergenic spots printed on the array are defective due to PCR failure.

Although we confirmed the presence of an epitope tag on each regulator, it is possible that DNA binding interactions may be lost or reduced due to the presence of the epitope tag. We previously noted that essentially identical results were obtained for specific regulators when immunoprecipitation is performed with epitope-tagged regulators or when it is performed with polyclonal antibodies against those regulators (4), so the epitope tag does not appear to modify binding interactions in these specific cases. We make the assumption that the lack of serious growth defects and the presence of positive binding results together indicate that the epitope modification does not alter binding activity substantially.

Rates of false negatives

The error model and the high confidence threshold selected to qualify as a positive result (p-value < 0.001) almost certainly play a major role in limiting genome coverage. The 0.001 p-value threshold was used to minimize false positive reports, with the expectation that this threshold may result in an underestimate of regulator-DNA interactions. If the p-value threshold is changed to 0.005 (a result where we would expect higher levels of false positives) 58% of yeast genes are bound by one of the 106 regulators. Our estimate of false negative rates (see QUALITY CONTROL section above) suggests that as many as 4300 out of the 6270 yeast genes (69%) may be bound by the 106 transcriptional regulators in yeast cells grown in rich medium.

Regulators that have yet to be identified

We have not subjected all of the transcriptional regulators encoded in the yeast genome to genome-wide location analysis. We used the Incyte Yeast BioKnowledge Library and the MIPS CYGD to identify all regulators with evidence for DNA binding activity and transcriptional regulatory activity. As more evidence is accumulated by investigators, the number of candidate regulators that would qualify using these criteria will increase. Thus, we might expect to cover an increasing percentage of the genome as we profile additional candidate regulators.

Contribution of extracellular environment

It is likely that a substantial proportion of genes that are not bound in the genome-wide dataset described here are activated under growth conditions other than the rich media condition used here. Yeast cells have evolved to live in a variety of challenging nutrient-limited environments. A change in the growth environment of yeast cells causes tremendous changes in its gene expression program, sometimes affecting the expression of as much as 1/3 of genes in the genome (5,6).

Among the 106 regulators profiled here in yeast cells grown in rich medium, the promoter regions for 2343 out of 6270 genes (37%) are bound by one or more regulators. For a subset of these regulators, we have assayed the effects of changing environmental conditions. A total of 35 regulators have been assayed in at least one additional growth condition. From a total of 53 experiments with regulators in these additional growth conditions, we find that 478 genes that are not bound by any regulator under rich media conditions are now bound by a regulator. Thus, we expect that further systematic studies of regulators under multiple conditions will more fully reveal the transcriptional regulatory network of the yeast cell.

COMPARISON BETWEEN LOCATION ANALYSIS DATA AND KNOWN SEQUENCE SPECIFICITY

Most of the transcription factors we have studied contain recognizable DNA-binding domains, and in several cases the binding sites for the factors have been characterized extensively. To test whether the binding specificities of factors in our assay were consistent with the published data, we searched for the expected binding site sequences in the bound intergenic regions. We found that many known binding site motifs are highly over-represented in the set of probes bound by the corresponding factor (Table S4). For example, the Cbf1 binding site (RTCACRTG) is present in 83% of the probes bound by this factor, but only 5% of all intergenic regions (p=9E-28). In cases where the enrichment is weaker, there may be interesting biological explanations. While recognition of specific sequences is an element of regulator binding, many other influences, including post-translational modifications, subcellular localization, interactions with other proteins and accessibility of the DNA, can affect which intergenic regions are bound by a particular regulator. For example, the Gcr1 binding site shows little enrichment in the probe set (p=0.02). Gcr1 functions in a complex with Gcr2 and Rap1, and its DNA-binding domain is dispensable for function (7), suggesting that Gcr1 binding can be mediated through multiple pathways. As a result, the presence of Gcr1 at intergenic regions *in vivo* may not necessarily correlate with the presence of the literature defined Gcr1 binding site.

The preferred binding sites for the majority of the 106 transcription factors studied are not known. However, the observed enrichment of known binding sites in bound probes suggests that it may be possible to infer these specificities from the location analysis data. We searched for sequences that are over-represented in the probes bound by each regulator using the program MEME (8). The discovered sequence motifs were scored for significance by two criteria: an E-value calculated by MEME and a specificity score described in (9). The derived motifs for all 106 activators are shown in Table S5 and the details of the methods are available at http://web.wi.mit.edu/young/regulator_network. MEME found the expected binding site for 13 of the 20 activators (65%) with well-characterized, compact binding sites in the TRANSFAC database (10) (Table S6).

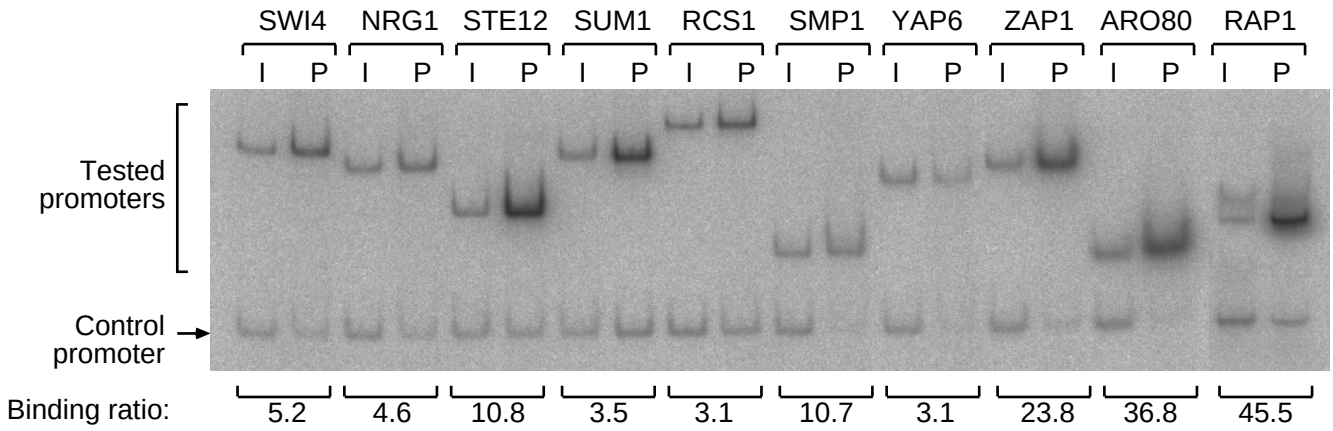
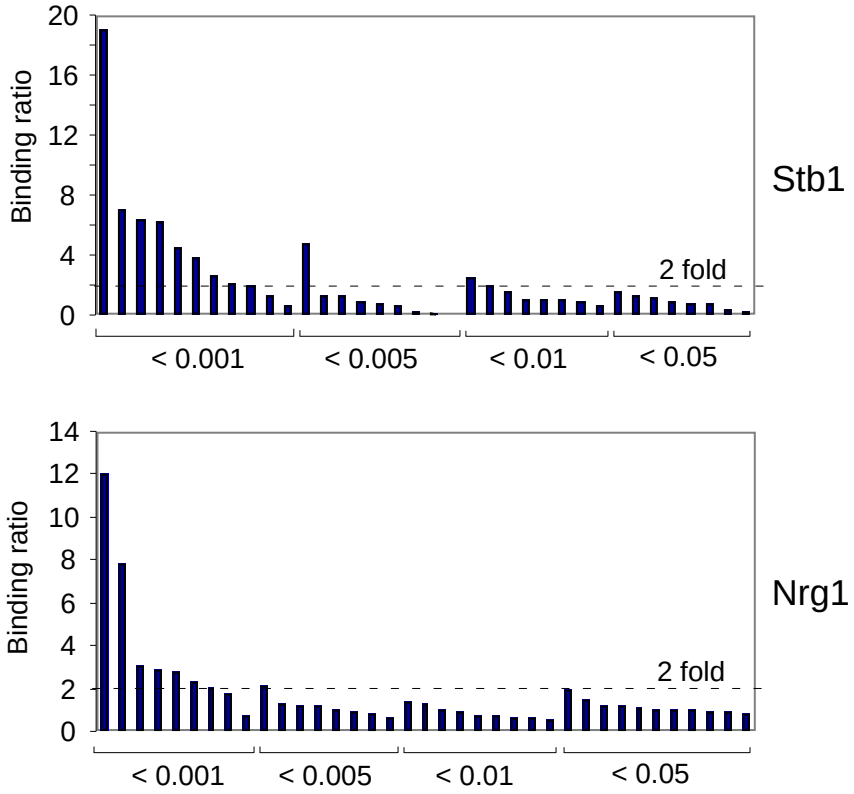


Figure S1.
Gene-specific PCR analysis of predicted regulator-gene interactions.

Shown are 10 examples of confirmation of location analysis results through gene-specific PCR analysis. The 10 regulators selected are examined here for their ability to bind their own promoters, interactions predicted using the 0.001 p-value threshold. The names of the transcriptional regulators are listed across the top. Primers designed to amplify their respective promoter were used (Tested promoters) and primers for the promoter of *ARN1* (Control promoter) were used as a negative control. The signals obtained from the input extracts (I) as well as the immunoprecipitated extracts (P) are shown and the binding ratio calculated as the enrichment of P over I after normalization with the control promoter are shown at the bottom (Binding ratio).

A



B

	p-value thresholds			
Factor	**< 0.001**	**< 0.005**	**< 0.01**	**< 0.05**
Stb1	9/11	1/9	2/8	0/8
Nrg1	7/9	1/8	0/9	0/11

Figure S2.
Estimation of false positive and false negative rates by chromatin immunoprecipitations (ChIP) and gene-specific PCR analysis.

A) For each regulator, Stb1 (top panel) and Nrg1 (bottom panel), we performed gene-specific PCR analysis on sets of genes predicted to bind the respective regulator using four different p-value thresholds. The data for each regulator-gene interaction is shown as a bar where the height represents the fold enrichment (Binding ratio as calculated above). The data within each panel is grouped according to which p-value threshold was used to predict the regulator-gene interaction. A 2-fold binding ratio (indicated by the dashed line) was used as the cutoff for the PCR analysis.

B) Summary of results from part A. We show the number of confirmed interactions out of the number of interactions tested for each regulator at a particular p-value threshold. Regulators are listed on the left. Thresholds are listed across the top.

Table S1. Regulator-gene interactions confirmed by gene specific PCR

Factors	Genes
Ace2	*SIC1, CTS1*
Aro80	*ARO80*
Cup9	*PTR2*
Dig1	*FUS1*
Fhl1	*RPS24B, RPL11A, RPS2, RPL19B, GCN4, FKH2*
Fkh1	*SWI5*
Fkh2	*SWI5, CDC20, CLN3*
Gal4	*GAL2, GAL3, GAL1/10, GAL7, MTH1, FUR4, PCL10, GAL80, PGM2, GCY1*
Gcn4	*LEU3, MET4, RTG3, UGA3, PUT3*
Mbp1	*TMP1*
Mcm1	*SWI5, STE2, CLB1, CDC20, CLN3*
Msn4	*ROX1, YAP1, RPN4*
Ndd1	*SWI5*
Nrg1	*HXT16, YIR013C, RPI1, YIR020C, SOR1, HXT15, SGA1, NRG1*
Pho2	*PHO5, PHO84*
Pho4	*PHO5, PHO84*
Rap1	*RAP1, YOL082W, RPL11A, TYE7, MSN4, BDF1, YRF1-6, HSP12, RPS2, RPL2A, HIS4*
Rcs1	*RCS1*
Rme1	*CLN2*
Rtg3	*IDH2, ACO1*
Smp1	*SMP1*
Stb1	*YDR501W, YOR246C, YOR138C, NDD1, GIC1, YKL097C, FKS1, PCL1, CLN2*
Ste12	*FUS1, STE2, FAR1, FUS1, STE2, FAR1, STE12*
Sum1	*SUM1*
Swi4	*HO, TMP1, SWI4*
Swi6	*TMP1*
Thi2	*THI80, THI7*
Yap6	*YAP6*
Zap1	*ZAP1*

Regulator-gene interactions were predicted by genome-wide location analysis under varied experimental conditions using a 0.001 p-value threshold. Genes listed twice for Ste12 indicate Ste12 binds those genes under two different experimental conditions tested.

Table S2. Comparison between location analysis and published literature

Factors	Genes	p-value	Factors	Genes	p-value
Ace2	*CTS1*	4.30E-09	Swi4	*HO*	4.08E-07
	EGT2	4.58E-06		*MNN1*	1.87E-06
				OCH1	3.71E-06
Fkh1	*SWI5*	3.33E-02		*FKS1*	6.00E-04
	CLB2	3.15E-03		*CLN1*	1.24E-06
				GAS1	4.90E-09
Fkh2	*SWI5*	6.26E-07		*PCL1*	8.94E-09
	YJL051W	6.22E-08		*CLN2*	5.02E-03
	ACE2	2.22E-09		*KRE6*	3.65E-03
	CLB2	6.81E-08			
			Swi6	*HO*	2.37E-07
Mbp1	*RNR1*	2.78E-08		*RNR1*	7.16E-13
	CDC6	1.78E-04		*SWI4*	1.31E-06
	CDC21	8.94E-07		*CDC6*	2.29E-03
	CLB5	1.06E-04		*CLN1*	2.18E-05
				PCL1	9.20E-05
Ndd1	*SWI5*	1.55E-15		*CDC21*	4.41E-06
	CLB2	0.00E0		*CLN2*	2.32E-02
				CLB5	3.70E-03
Mcm1	*SWI5*	3.22E-09			
	MFA1	4.41E-06	Swi5	*PCL2*	8.11E-04
	SWI4	7.89E-05		*PCL9*	1.70E-07
	STE2	5.46E-05		*ASH1*	3.49E-07
	FAR1	6.92E-03		*SIC1*	3.77E-05
	CDC6	5.68E-04		*CTS1*	1.61E-03
	STE6	6.55E-04		*EGT2*	1.03E-14
	ACE2	1.28E-07			
	CDC46	2.33E-05			
	MFA2	2.94E-04			
	CLB2	1.31E-05			
	CLN3	2.44E-05			

Regulator-gene interactions reported in the literature for nine cell cycle regulators (3) and the corresponding p-values from genome-wide location analysis data.

Table S3. Confirmed regulator-gene interactions predicted at different p-value thresholds

p-value ranges	Factors	Genes
< 0.001	Nrg1	*HXT16, YIR013C, RPI1, YIR020C, SOR1, HXT15, SGA1*
	Stb1	*YDR501W, YOR246C, YOR138C, NDD1, GIC1, YKL097C, FKS1, PCL1, CLN2*
< 0.005	Nrg1	*ENA1*
	Stb1	*HCM1*
< 0.01	Stb1	*SPT21, SRO4*

Listed above are the regulator-gene interactions displayed in Figure S2. The predicted interactions were selected using various p-value thresholds (indicated on the left) and confirmed by chromatin immunoprecipitation and subsequent gene-specific PCR analysis.

Table S4. Enrichment of known sequence motifs in intergenic regions bound by regulators

Regulator	Known Consensus Motif	Frequency in Bound Probes (%)	Frequency in All Intergenic Regions (%)	Enrichment P-value
Abf1	TCRNNNNNNACG	96	24	8E-101
Cbf1	RTCACRTG	83	5	1E-27
Gal4	CGGNNNNNNNNNNCCG	12	2	0.001
Gcn4	TGACTCA	62	3	5E-39
Gcr1	CTTCC	71	51	0.02
Hap2	CCAATNA	58	21	9E-05
Hap3	CCAATNA	58	21	6E-05
Hap4	CCAATNA	60	21	8E-10
Hsf1	GAANNTTCNNGAA	43	1	6E-21
Ino2	ATGTGAAA	56	4	7E-06
Mata(A1)	TGATGTANNT	15	3	0.06
Mcm1	CCNNNWWRGG	64	17	3E-18
Mig1	WWWWSYGGGG	0	3	1
Pho4	CACGTG	10	5	0.1
Rap1	RMACCCANNCAYY	20	1	3E-24
Reb1	CGGGTRR	81	12	6E-54
Ste12	TGAAACA	45	11	3E-13
Swi4	CACGAAA	24	7	2E-09
Swi6	CACGAAA	27	7	3E-08
Yap1	TTACTAA	54	12	9E-11

The regulators listed above have high quality binding data in the TRANSFAC database (TRANSFAC assigned “quality” value of 3 or better). For each regulator, we identified a consensus binding motif that incorporated data from both TRANSFAC and the literature. We then calculated the frequency of the motif in the set of intergenic regions bound by the relevant factor and the frequency of the motif in all intergenic regions. The p-value represents the probability of randomly selecting a set of intergenic regions that contains the consensus binding motif at, or above, the frequency observed in the experimental data and were calculated from a hypergeometric distribution.

Table S5: Putative DNA binding motifs determined by MEME.

Regulator	Best Motif (scored by E-value)	Best Motif (scored by Hypergeometric)
Abf1	TYCGT--R-ARTGAYA	TYCGT--R-ARTGAYA
Ace2	RRRAARARAA-A-RARAA	GTGTGTGTGTGTGTG
Adr1	A-AG-GAGAGAG-GGCAG	YTSTYSTT-TTGYTWTT
Arg80	T--CCW-TTTKTTTC	GCATGACCATCCACG
Arg81	AAAAARARAAAARMA	GSGAYARMGGAMAAAAA
Aro80	YKYTYTTYTT----KY	TRCCGAGRYW-SSSGCGS
Ash1	CGTCCGGCGC	CGTCCGGCGC
Azf1	GAAAAAGMAAAAAAA	AARWTSGARG-A--CSAA
Bas1	TTTTYYTTYTTKY-TY-T	CS-CCAATGK--CS
Cad1	CATKYTTTTTTKYTY	GCT-ACTAAT-
Cbf1	CACGTGACYA	CACGTGACYA
Cha4	CA---ACACASA-A	CAYAMRTGY-C
Cin5	none	none
Crz1	GG-A-A--AR-ARGGC-	TSGYGRGASA
Cup9	TTTKYTKTTY-YTTTKTY	K-C-C---SCGCTACKGC
Dal81	WTTKTTTTTYTTTTT-T	SR-GGCMCGGC-SSG
Dal82	TTKTTTTYTTC	TACYACA-CACAWGA
Dig1	AAA--RAA-GARRAA-AR	CCYTG-AYTTCW-CTTC-
Dot6	GTGMAK-MGRA-G-G	GTGMAK-MGRA-G-G
Fhl1	-TTWACAYCCRTACAY-Y	-TTWACAYCCRTACAY-Y
Fkh1	TTT-CTTTKYTT-YTTTT	AAW-RTAAAYARG-
Fkh2	AAARA-RAAA-AAAR-AA	GG-AAWA-GTAAACAA
Fzf1	CACACACACACACACAC	SASTKCWCTCKTCGT
Gal4	TTGCTTGAACGSATGCCA	TTGCTTGAACGSATGCCA
Gal4 (Gal)*	YCTTTTTTTTYTTYYKG	CGGM---CW-Y--CCCG
Gat1	none	none
Gat3	RRSCCGMCGMGRCGCGCS	RGARGTSACGCAKRTTCT
Gcn4	AAA-ARAR-RAAAARRAR	TGAGTCAY
Gcr1	GGAAGCTGAAACGYMWRR	GGAAGCTGAAACGYMWRR
Gcr2	GGAGAGGCATGATGGGGG	AGGTGATGGAGTGCTCAG
Gln3	CT-CCTTTCT	GKCTRR-RGGAGA-GM
Grf10	GAAARRAAAAAAMRMARA	-GGGSG-T-SYGT-CGA
Gts1	G-GCCRS--TM	AG-AWGTTTTTGWCAAMA
Haa1	none	none
Hal9	TTTTTTYTTTTY-KTTTT	KCKSGCAGGCWTTKYTCT
Hap2	YTTCTTTTYT-Y-C-KT-	G-CCSART-GC
Hap3	T-SYKCTTTTCYTTY	SGCGMGGG--CC-GACCG
Hap4	STT-YTTTY-TTYTYYYY	YCT-ATTSG-C-GS
Hap5	YK-TTTWYYTC	T-TTSMTT-YTTTCCK-C
Hir1	AAAA-A-AARAR-AG	CCACKTKSGSCCT-S
Hir2	WAAAAAAGAAAA-AAAAR	CRSGCYWGKGC
Hms1	AAA-GG-ARAM-AARAA	GC-GGGCAC-C
Hsf1	TYTTCYAGAA--TTCY	TYTTCYAGAA--TTCY
Ime4	CACACACACACACACACA	CACACACACACACACACA
Ino2	TTTYCACATGC	SCKKCGCKSTSSTTYAA
Ino4	G--GCATGTGAAAA	G--GCATGTGAAAA
Ixr1	GAAAA-AAAAAAAARA-A	CTTTTTTTYYTSGCC
Leu3	GAAAAARAARAA-AA	GCCGGTMMCGSYC--
Mac1	YTTKT--TTTTTYTYTTT	A--TTTTTYTTKYGC
Mal13	GCAG-GCAGG	AAAC-TTTATA-ATACA
Mal33	none	none

Mata1	GCCC-C	CAAT-TCT-CK
Mbp1	TTTYTYKTTT-YYTTTTT	G-RR-A-ACGCGT-R
Mcm1	TTTCC-AAW-RGGAAA	TTTCC-AAW-RGGAAA
Met31	YTTYYTTYTTTTYTYTTC	GCACGTGATS
Met4	MTTTTTYTYTYTTC	S-CCACAGTTKT
Mig1	TATACA-AGMKRTATATG	KAASARA-YTCGAGYSC
Mot3	TMTTT-TY-CTT-TTTWK	TG-AAGCA-YAGRTT-TT
Msn1	KT--TTWTTATTCC-C	KT--TTWTTATTCC-C
Msn2	ACCACC	GTAGAAAA--CAT-CC
Msn4	R--AAAA-RA-AARAAAT	AAG-G-AGSCSCGG-TGG
Mss11	TTTTTTTTCWCTTTKYC	GTTC-TTTGCTG-TR
Ndd1	TTTY-YTKTTTY-YTTYT	GG-AAWW-GTMAACA
Nrg1	TTY--TTYTT-YTTTYYY	SC--YSCC-CCSC-GCS
Pdr1	T-YGTGKRYGT-YG	ACCCTG-SCCATT-ACCC
Phd1	TTYYYTTTTTYTTTTYTT	TK-M-YGAARTAWSRGCM
Pho4	GAMAAAAAARAAAAR	TTGTACACTTYGTTT
Put3	CYCGGGAAGCSAMM-CCG	CYCGGGAAGCSAMM-CCG
Rap1	GRTGYAYGGRTGY	GRTGYAYGGRTGY
Rcs1	KMAARAAAAARAAR	GKGTGTGTGYGTGTGTGT
Reb1	RTTACCCGS	RTTACCCGS
Rfx1	AYGRAAAARARAAAARAA	CGTTRCCATGGCAACC
Rgm1	GGAKSCC-TTTY-GMRTA	AKACASACR-WCTTMC-C
Rgt1	CCCTCC	CCCTCC
Rim101	GCGCCGC	GCGCCGC
Rlm1	TTTTC-KTTTYTTTTTC	A-CTSGAAGAAATGCGGT
Rme1	ARAAGMAGAAARRAA	TGWSATGGMGRAAGS
Rox1	YTTTTCTTTTY-TTTTT	TMS-CTSCTGCAGGS
Rph1	ARRARAAAGG-	CAATGASGWSYAKAYCTS
Rtg1	YST-YK-TYTT-CTCCCM	YST-YK-TYTT-CTCCCM
Rtg3	GARA-AAAAR-RAARAAA	C-G-TTTT-TTGGCC
Sfl1	CY--GGSSA-C	CAMAKA-SK-TG--ATGA
Sfp1	CACACACACACACAYA	CACACACACACACAYA
Sip4	CTTYTWTTKTTKTSA	AGC-GGCAT-T--C
Skn7	YTTYYYTYTTTYTYYTTT	G--S-GGGCYS-SC
Sko1	none	none
Smp1	AMAAAAARAARWARA-AA	CACACACACACMCACAYA
Sok2	ARAAAARRAAAAAG-RAA	GC-SS-YATSYTTTTCTT
Stb1	RAARAAAAARCMRSRAAA	SRSSRMTTTTSGCGT
Ste12	TTYTKTYTY-TYYKTTTY	GSAASRR-TGATRAWGYA
Stp1	GAAAAMAA-AAAAA-AAA	CCGTACGGCG--GC
Stp2	YAA-ARAARAAAAA-AAM	TTGACGTKRTT
Sum1	TY-TTTTTTYTTTTT-TK	CCCGWCGCGWCGCGS
Swi4	RAARAARAAA-AA-R-AA	CSMRRCGCGAAAA
Swi5	CACACACACACACACACA	CACACACACACACACACA
Swi6	RAARRRAAAAA-AAAMAA	WCGCGTCGCGTY-C
Thi2	GCCAGACCTAC	CT-T-TC-C-AATAT--C
Uga3	GG-GGCT	RCASWGCYTRAWAC-TGC
Yap1	TTYTTYTTYTTTY-YTYT	TGMTKACTAAT
Yap3	none	none
Yap5	YKSGCGCGYCKCGKCGGS	GGATGCCMTTTYMGAATA
Yap6	TTTTYYTTTTYYYYKTT	GCCTGY-CMTTSSCCSGY
Yap7	none	none
Yfl044c	TTCTTKTYYTTTT	MSMAAA-GRAARRWA-G
Yjl206c	TTYTTTTYTYYTTTYTTT	T-C-TTYYY-GTRM-GMA
Zap1	TTGCTTGAACGGATGCCA	AGCACCSRCATGAGT-TG
Zms1	MG-MCAAAAATAAAAS	MG-MCAAAAATAAAAS

We searched for potential regulator binding motifs by using the program MEME to search the intergenic regions bound by regulators for overrepresented sequences. For each regulator, the sequences of intergenic regions bound with p-values less than 0.001 were extracted to use as input for motif discovery. MEME was run using the following settings: a motif width ranging from 6 to 18 bases, the "zoops" distribution model, a 6th order Markov background model and a discovery limit of 20 motifs. The discovered sequence motifs were scored for significance by two criteria: an E-value calculated by MEME and a specificity score described in (9). The motif with the best score using each metric is shown for each regulator. All motifs presented are derived from datasets generated in rich growth conditions with the exception of a previously published dataset for epitope-tagged Gal4 grown in galactose (4) (indicated on the table by an asterisk).

Table S6: Comparison of the DNA binding motifs identified by MEME and the motifs compiled in the TRANSFAC database.

Regulator	Match in TRANSFAC	MEME Motif	Matching Transfac Entry	Rank of matching motif (MEME E-value)	Rank of matching motif (Hypergeometric)
Abf1	YES	TYCGT--R-ARTGAYA	AAGTGCCGTGCATAATGATGTGGG	1	1
Cbf1	YES	CACGTGACYA	RTCACRTG	1	1
Gal4	NO	n/a	n/a	n/a	n/a
Gal4 (Gal) *	YES	CGGM---CW-Y--CCCG	CGGACCACAGTACTCCG	2	1
Gcn4	YES	TGAGTCAY	SNSNNNNNRTGACTCATNS	2	1
Gcr1	YES	GGAAGCTGAAACGYMWRR	RGCTTCCWC	1	1
Hap2	NO	n/a	n/a	n/a	n/a
Hap3	NO	n/a	n/a	n/a	n/a
Hap4	YES	YCT-ATTSG-C-GS	YCNNCCAATNA	3	1
Hsf1	YES	TYTTCYAGAA--TTCY	CTaGAAcgTTCTAGAAgcTTCgAG	1	1
Ino2	YES	TTTYCACATGC	ATGTGAAAA	1	2
Mata1	NO	n/a	n/a	n/a	n/a
Mcm1	YES	TTTCC-AAW-RGGAAA	WTWCCYAAWNNGGTAA	1	1
Mig1	NO	n/a	n/a	n/a	n/a
Pho4	NO	n/a	n/a	n/a	n/a
Rap1	YES	GRTGYAYGGRTGY	ACACCCATACATCT	1	1
Reb1	YES	RTTACCCGS	YNNYYACCCG	1	1
Ste12	NO	n/a	n/a	n/a	n/a
Swi4	YES	CSMRRCGCGAAAA	CGCGAAA	2	1
Swi6	YES	ACGCGARA-GCSAA	CGCGAAA	3	2
Yap1	YES	TGMTKACTAAT	CTGACTAA	2	1

We compared the results derived from MEME analysis to experimental results in the TRANSFAC database for 20 regulators that had compact binding motifs with a TRANSFAC-assigned “quality” value of 3 or better. For each regulator, MEME generated candidate binding motifs based on overrepresentation of the motifs in the set of intergenic regions bound by the regulator. Motifs were converted to position-specific scoring matrices (PSSMs) and used to scan a database of TRANSFAC entries. A motif was considered a match if the PSSM score of a TRANSFAC entry was at least 70% of the optimal score for the PSSM. In cases where there was a match, we show the motif identified by MEME and the matching TRANSFAC entry. Sections of the motifs that are underlined represent regions with the highest similarity (note some matches align to the reverse complement of TRANSFAC entries). The rank columns indicate the position of the motif on the list of the top 20 candidate motifs as scored by each of two metrics as shown. All motifs presented are derived from datasets generated in rich growth conditions with the exception of a previously published dataset for epitope-tagged Gal4 grown in galactose (4) (indicated on the table by an asterisk).

REFERENCES

1. M. P. Cosma, et al., *Cell* **97**, 299-311. (1999).
2. T. R. Hughes, et al., *Cell* **102**, 109-26. (2000).
3. I. Simon, et al., *Cell* **106**, 697-708. (2001).
4. B. Ren, et al., *Science* **290**, 2306-9. (2000).
5. H. C. Causton, et al., *Mol Biol Cell* **12**, 323-37. (2001).
6. A. P. Gasch, et al., *Mol Biol Cell* **11**, 4241-57. (2000).
7. J. Tornow, et al., *EMBO* J. **12**, 2431-7. (1993).
8. T. Bailey, C. Elkan. *Proceeding of the Second International Conference on Intelligen Systems for Molecular Biology* (AAAI Press. Menlo Park, California, 1994). pp.28-36.
9. J. D. Hughes, et al., *J Mol Biol* **296**, 1205-14. (2000).
10. E. Wingender, et al., *Nucleic Acids Res.* **28**, 316-9