# Identifying transcription factor functions and targets by phenotypic activation

Gordon Chua*, Quaid D. Morris*†‡, Richelle Sopko§, Mark D. Robinson*‡, Owen Ryan§, Esther T. Chan§, Brendan J. Frey*‡, Brenda J. Andrews*§, Charles Boone*§, and Timothy R. Hughes*§¶

*Banting and Best Department of Medical Research, and Departments of †Computer Science, ‡Electrical and Computer Engineering, and §Medical Genetics and Microbiology, University of Toronto, 160 College Street, Toronto, ON, Canada M5S 1A8

Communicated by Stanley Fields, University of Washington, Seattle, WA, June 19, 2006 (received for review March 14, 2006)

**Mapping transcriptional regulatory networks is difficult because many transcription factors (TFs) are activated only under specific conditions. We describe a generic strategy for identifying genes and pathways induced by individual TFs that does not require knowledge of their normal activation cues. Microarray analysis of 55 yeast TFs that caused a growth phenotype when overexpressed showed that the majority caused increased transcript levels of genes in specific physiological categories, suggesting a mechanism for growth inhibition. Induced genes typically included established targets and genes with consensus promoter motifs, if known, indicating that these data are useful for identifying potential new target genes and binding sites. We identified the sequence 5′-TCACGCAA as a binding sequence for Hms1p, a TF that positively regulates pseudohyphal growth and previously had no known motif. The general strategy outlined here presents a straightforward approach to discovery of TF activities and mapping targets that could be adapted to any organism with transgenic technology.**

microarray | overexpression | yeast

Delineation of transcriptional control networks is critical to understanding how the physiology of cells and organisms is orchestrated. One of the most surprising results of genome sequencing from yeast to vertebrates is the large amount of conserved intergenic sequence, much of which is presumably cis-regulatory (1–3). Moreover, in most sequenced genomes, a correspondingly large proportion of genes appear to encode transcription factors (TFs), typically 3–6% of all genes (4, 5). Even in yeast, a relatively well studied organism, physiological functions and/or DNA-binding sites remain unknown for roughly half of all apparent sequence-specific DNA-binding TFs (4, 6), suggesting that there are many more transcriptional regulatory pathways than are currently known.

Several strategies have been devised to decipher regulatory codes, but none is without caveats. Algorithms that seek conserved promoter elements (1, 2) or common sequence elements in promoters of coexpressed genes (7, 8) can identify potential cis-regulatory sequences, but do not inherently identify the binding TF. Microarray-based biochemical approaches promise to rapidly identify sequence preferences of individual TFs, but additional influences apparently contribute to site occupancy *in vivo* (9, 10). ChIP-chip (4, 11, 12) identifies sequences bound by a TF *in vivo*, but positive results often depend on identifying conditions under which the TF is DNA-bound; moreover, bound sites may not be active (13).

Artificial activation of TFs by genetic modification is a promising experimental strategy for demonstrating functionality of TFs *in vivo* without knowing the natural condition under which the TF acts. Devaux *et al.* (14) showed a nearly perfect correspondence between the target genes activated by a well studied gain-of-function mutation in *PDR1* (*PDR1-3*), and those activated by an inducible fusion protein consisting of the Pdr1p DNA-binding domain (DBD) and the Gal4p activation domain. Other studies have examined the effects of overexpressing native TFs (15–17). However, to our knowledge, this general approach has not yet been tested on a large scale to ask whether it is generally effective in specifically activating primary targets of TFs, or whether there is any way to determine which TFs are likely to be amenable to this type of experimentation.

In a systematic genetic screen using an ordered clone set overexpressing full-length ORFs from the GAL promoter (18) we found that 57 of 175 yeast TFs tested (32.6%) caused growth inhibition when overexpressed. This number is more than twice as many as would be expected by chance: over the entire genome, we found that only 769 of 5,280 (14.6%) of genes caused growth inhibition, and in fact TFs are among the functional classes that are most toxic when overexpressed (18). This finding suggested that in many cases a TF might be activated by simple overexpression, even if the TF is not normally active under the specific growth condition used. To ask whether this is the case, and, if so, whether the resulting transcription profiles reflected known or apparent physiological functions of the TFs, we have now analyzed these TF overexpression strains by using DNA microarrays. Here, we show that in many cases the induced genes correspond to physiological functions and known targets and that expected binding sites of the TF can usually be identified in the promoters of these genes. Markedly fewer expression changes were observed in deletion mutants of this same collection of TFs, consistent with the view that specific regulatory events or conditions are prerequisites for activation of many TFs. We demonstrate that the basic helix–loop–helix family member Hms1p (19) binds *in vitro* to a cis-regulatory sequence predicted from the overexpression data and that overexpression of two of the apparent target genes causes the same pseudohyphal growth phenotype displayed by cells overexpressing *HMS1*. Together, these results suggest that analysis of gene expression in organisms in which TF overexpression causes a visible phenotype, a phenomenon we term “phenotypic activation,” represents a straightforward approach for rapidly characterizing TFs and mapping regulatory networks on a large scale.

## Results

**Overexpression of TFs Results in Diverse and Dramatic Transcriptional Responses.** Our previous analysis (18) identified 57 TFs that caused growth inhibition when overexpressed. An initial two-color microarray expression analysis of one of these, GAL-*GCN4* (compared with the empty vector control; Fig. 4, which is published as supporting information on the PNAS web site), showed that many of the induced genes were known physiological targets of Gcn4p (20) and that virtually all of the catalogued Gcn4p targets were induced (see below). Gcn4p is a well characterized example of a TF whose deletion is phenotypically

Conflict of interest statement: No conflicts declared.

Freely available online through the PNAS open access option.

Abbreviations: TF, transcription factor; WMW, Wilcoxon–Mann–Whitney; DBD, DNA-binding domain; MBP, maltose-binding protein.

¶To whom correspondence should be addressed. E-mail: t.hughes@utoronto.ca.

benign except in specific circumstances; nearly all genes encoding amino acid biosynthetic enzymes are induced by Gcn4p in amino acid-deprived cells (21). Our observation that overexpression of *GCN4* was sufficient to induce a physiologically relevant response suggests that simple mass action may produce a relatively "natural" hyperactivation state and that the growth inhibition may be caused by inappropriate induction of the biosynthetic pathways controlled by Gcn4p. Moreover, Gcn4p DNA-binding sequences were enriched specifically among genes with the highest ratios (Fig. 5, which is published as supporting information on the PNAS web site) and the known Gcn4p DNA-binding site was perfectly recapitulated by seeking sequence motifs that correlated with the degree of gene induction (see below).

To ask whether Gcn4p is an exception, and whether overexpressing different TFs resulted in different transcriptional responses, we analyzed a total of 55 TF overexpression strains by using microarrays (the mating-type determinants *MATα2* and *HMRa2* were omitted) and corresponding deletion mutants of 51 nonessential TFs for comparison (grown under a single standard condition). A fluor-reversal strategy was used in which mRNA from each strain was compared with mRNA from an empty-vector control (in the case of overexpression strains) or WT strain (in the case of deletion mutants) twice, with the red/green fluors reversed in the replicate. Each strain was examined at a single 3-h time point, as a time course of *GCN4* induction indicated that little information is gained from taking additional time points (Fig. 6, which is published as supporting information on the PNAS web site, and data not shown).

To isolate experiments in which the transcriptional alterations could not be accounted for by measurement noise or effects caused by slow growth, we identified those in which (*i*) the replicates had a Pearson correlation >0.3, which typically separates physiologically unrelated experiments (22), or (*ii*) the fluor-reversal experiments have reciprocal best-matching correlations among all dye swaps and vice versa. Forty-six TF overexpression experiments passed at least one of these criteria, indicating that the vast majority of TF overexpression microarray data contained distinctive and prominent patterns. This finding is illustrated in Fig. 1*A*, which shows all genes induced in any of the 46 overexpression experiments. In contrast, only 10 of the TF deletion microarray experiments passed these criteria (all of which were TFs represented among the 46 passing overexpressors) largely because there were few expression changes in these mutants beyond measurement noise, such that few experiments contained a distinctive pattern. Fig. 1*B* illustrates that there is less expression change in deletion mutants versus overexpressed TFs and also suggests that there is little correspondence between the genes induced upon overexpression and those whose expression is reduced in the deletion mutant (Fig. 1 *A* and *B*). Thus, it is possible that many TFs are inactive under typical unstressed growth conditions, which could account for the fact that it has been difficult to obtain meaningful ChIP-chip data for roughly half of all apparent yeast TFs (4).

**Overexpression of TFs Induces or Represses Known Targets and Pathways.** Three lines of evidence indicate that genes induced in these experiments are likely to be physiological targets. First, most of the TF overexpression experiments displayed specific and significant induction of genes in one or more Gene Ontology categories, using the Wilcoxon–Mann–Whitney (WMW) test, which calculates a *P* value (Fig. 1*C*) for differences in the median expression ratio ranks between genes that are in a given category and those that are not. In many cases, the significant categories were related to the known specific functions of the TF. For example, whereas amino acid biosynthesis categories were induced by overexpression of *GCN4*, overexpression of *UPC2* or *ECM22* (23) resulted in a general induction of genes in the ergosterol biosynthetic pathway (Fig. 1*C*). We obtained similar results for known repressors (e.g., *ROX1*), which are much fewer in number in our study (data not shown). These trends were readily distinguished even when the experiments also contained common transcriptional alterations characteristic of growth inhibition such as induction of stress-response genes and reduction of protein biosynthesis genes; these are visible as horizontal red and green bands in Fig. 1*A*.

Second, among the transcriptional activators and repressors we analyzed, and for which known target genes are present in TRANSFAC (24), we generally observed induction or repression of appropriate targets. Fig. 2 shows a comparison of WMW *P* values obtained for TRANSFAC targets for our overexpression data and "ChIP-chip" experiments done with these same TFs (4) (Fig. 2). As above, these tests measure how well the known targets are sorted to the top of the ranked list of genes. In most cases, overexpression yielded more significant discrimination of known targets than ChIP-chip by this test. For example, the three known Adr1p targets (*ACS1*, *CTA1*, and *ADH2*) have significantly higher ranks among induced genes in our data (7, 13, and 15 of 5,222), in comparison to their ranks in ChIP-chip data (1,359, 2,510, and 3,148 of 6,229). Cases where ChIP-chip yields greater significance may represent instances where overexpression does not result in induction of physiological targets; Ino2p is likely such an example. However, others may involve sampling artifacts: Met4p has only two targets in TRANSFAC but only one of them (*MET16*) is present in our final data set (where it is ranked 450 of 5,222).

Third, among the 25 TFs in our experiments with well known DNA-binding specificities, in 15 cases the established sites with at least a 75% match (i.e., 75% of the bases in the known motif were present in the found motif, without gaps) were identified in *de novo* motif searches, often as the top-scoring motif (Fig. 3). We initially ran a Gibbs sampling program (BioProspector) (25) on the highest 10, 30, and 50 scoring genes in each experiment; however, these analyses were often confounded by stress response elements (CCCCT) appearing in many of the induced genes, presumably as a secondary effect. We therefore developed a probabilistic inference algorithm called RankMotif (see *Methods*) that seeks both a motif specific to the individual experiment and a second motif that pervades multiple experiments. In addition to identifying known motifs, RankMotif generated high-scoring predicted binding sites for several TFs without established binding specificities. The full results are available on request. Fig. 3*A* shows the nine top-scoring transcriptional activators for which a binding specificity is known; in eight cases, we obtained at least a partial match (underlined in purple). Fig. 3*A* also shows motifs predicted for nine TFs for which there is no established binding specificity but for which the RankMotif *z*-score is comparable to the nine known activators shown.

The fact that known TF targets, expected functional categories, and known binding sites can be readily identified in these data indicates that there is a strong tendency for TF overexpression to cause meaningful transcriptional alterations. Although we cannot assume that all of the genes induced by overexpression of a TF are primary physiological targets (they might encompass both physiological and nonphysiological secondary effects and nonphysiological primary targets that are induced by overexpression) we reasoned that these data should facilitate identification of TF functions, target genes, and DNA-binding sites.

***HMS1* Overexpression Induces Pheromone-Responsive and Metabolic Genes, and Hms1p Binds 5′-TCACGCAA.** Figs. 1 and 3*A* contain undiscovered functions, targets, and binding sites for a variety of yeast TFs. Among the poorly characterized TFs for which overexpression yielded both induction of significant Gene On-

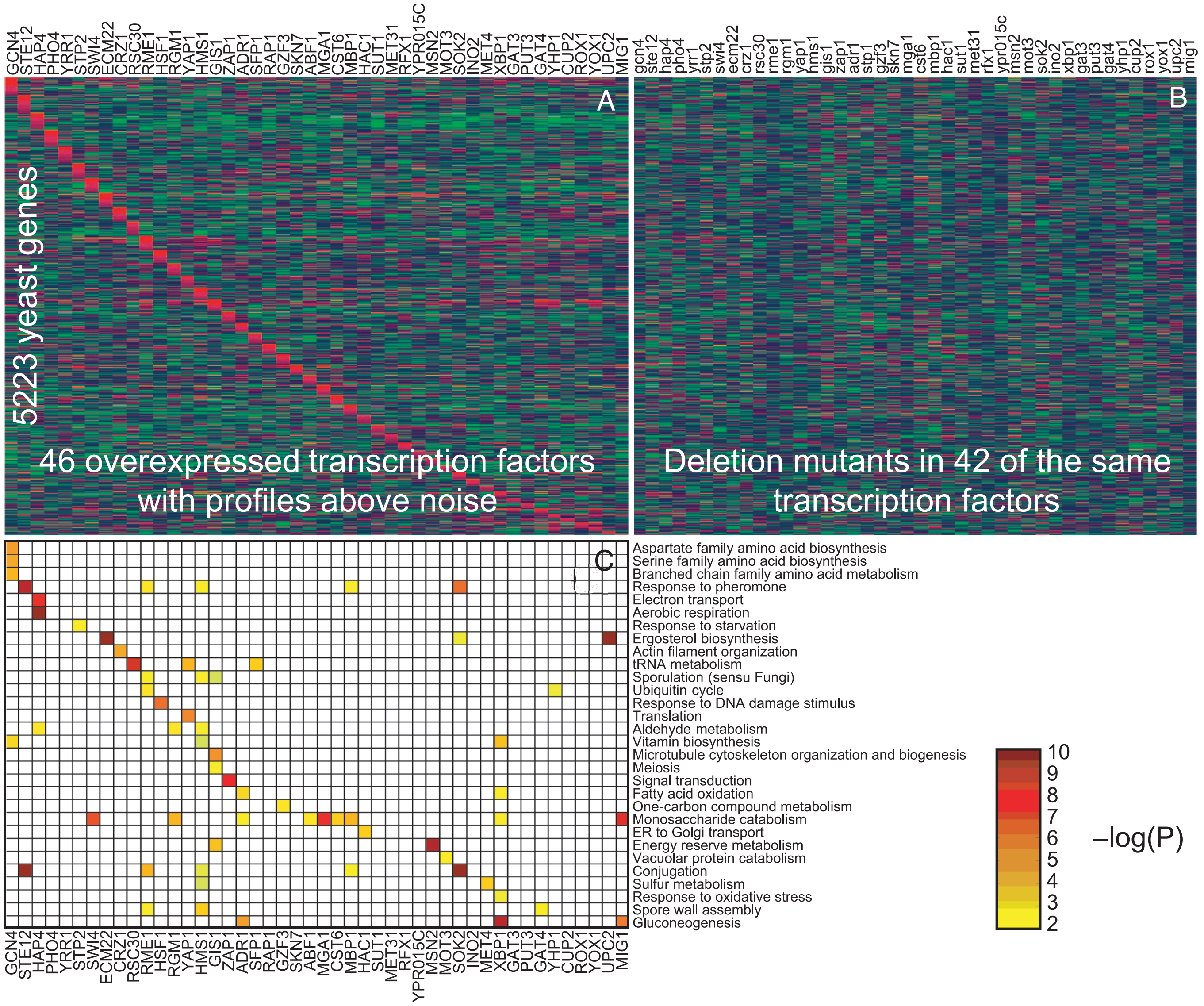

**Fig. 1.** Microarray expression data resulting from overexpression and/or deletion of 57 TFs that cause growth inhibition when overexpressed. Only TFs that contain expression profiles significantly above microarray noise when overproduced are shown. The diagram shows all 5,222 genes represented on the array after removal of dubious ORFs, transposable elements, mitochondria-encoded genes, and bad spots on the array. (*A*) Overexpression experiments. *z*-score-transformed data are shown (see *Methods*). Genes are ordered such that those with the greatest level of induction when a given TF is overexpressed are grouped, and then TFs are ordered according to the number of genes meeting this criterion. The color scale reflects *z*-score, which reflects noise-corrected log(ratio) (see *Methods*) and extends from ≈3-fold induction (red) to ≈3-fold reduction (green). (*B*) Microarray expression data (*z*-score transformed) of the corresponding deletion mutants. Rows and columns are in the same order as *A*, except that four essential TFs are missing. (*C*) Induction of specific functional classes of genes in response to TF overexpression. The columns are in the same order as *A*. Induction was scored with the WMW *P* value (see *Methods*).

tology categories and a predicted binding motif, and for which previous ChIP-chip experiments produced no readily interpretable results (4), was *HMS1* (high copy Mep suppressor). *HMS1* encodes a basic helix–loop–helix protein implicated in pseudohyphal growth because ectopic expression promotes filamentation and suppresses the pseudohyphal defect of the high-affinity ammonium permease-deficient Δ*mep2*/Δ*mep2* strain (19). However, the precise physiological role of Hms1p remains obscure: there are no known target genes or pathways of transcriptional activation by Hms1p, and no Hms1p DNA-binding sites have been identified either *in vivo* or *in vitro*. In our microarray data, *HMS1* overexpression induced some of the same genes induced by *STE12* in response to pheromone (17) and genes in a variety of metabolic pathways (Fig. 7*A*, which is published as supporting information on the PNAS web site), either of which could provide a potential mechanism for its morphological effect: *STE12* is required for pseudohyphal growth (26) and nutritional cues stimulate filamentous growth (27).

Our data also led to a predicted binding consensus for Hms1p. We performed gel-shift assays with purified Hms1p DBD on specific sequences corresponding to some of the top-scoring degenerate motifs identified by RankMotif (Fig. 3*A*) and detected strong binding to 5′-TCACGCAA (Fig. 3*B*), which overlaps six of the bases shown in Fig. 3*A*. Binding of Hms1p-DBD to the 5′-TCACGCAA motif is specific, as we observed no binding of Hms1p-DBD to the consensus motifs of Gcn4p and Upc2p (Fig. 3*B*) nor to other sequences tested, including some other variants of the consensus (data not shown and Fig. 8, which is published as supporting information on the PNAS web site). We then examined whether genes that are up-regulated in

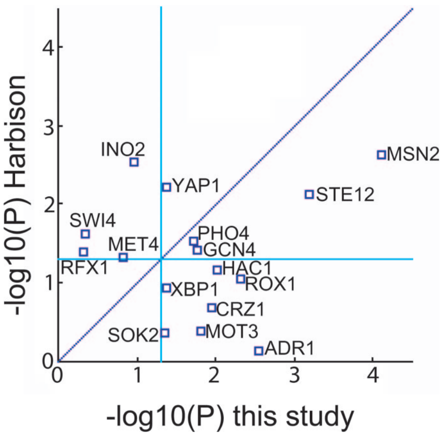

**Fig. 2.** Behavior of known TF targets in response to overexpression or deletion of the TF and compared with a similar analysis of genomewide ChIP-chip data from Harbison *et al.* (4). Each point indicates, for the TF indicated, the WMW *P* value (see *Methods*) for the difference of medians between the ranked TRANSFAC targets and those of all other ORFs; i.e., a point with a higher $-\log(P)$ value indicates that the median of TRANSFAC targets is shifted higher toward the top of the ranked list of genes. For our data, the *z*-scores are ranked; for Harbison *et al.* data, the *P* values are ranked. Only TFs with $P < 0.05$ in either Harbison *et al.* or this study are shown.

response to *HMS1* overexpression and contain exactly the 5′-TCACGCAA motif have a role in promoting pseudohyphal growth in a WT Σ1278 strain. We found that overexpression of either *URA10* or *YPC1*, which encode an orotate phosphoribosyltransferase and alkaline ceramidase, respectively (28, 29), promotes pseudohyphal growth and suppresses the pseudohyphal defect of the Δ*mep2*/Δ*mep2* diploid strain (Fig. 7*B*), although neither *URA10* or *YPC1* is by itself required for the *HMS1* hyperfilamentation phenotype (Fig. 7*C*). Intriguingly, there is evidence that both uracil biosynthesis and sphingolipid content impact pathogenesis and/or filamentation in pathogenic yeasts (30–33).

## Discussion

Our results show that phenotypic activation of TFs is feasible as a general approach to identifying TF activities, targets, and binding sites. Although further experimentation of individual cases will be required to conclusively distinguish all primary and secondary effects, the simple transient overexpression applied here yielded unique and meaningful results for the majority of TFs analyzed and these could be interpreted by objective statistical and machine learning techniques. Importantly, this approach appears to be much more fruitful than analysis of deletion mutants, possibly because most TFs are not active under typical growth conditions. Moreover, our results with Hms1p and other TFs (Fig. 3*B*) indicate that the approach also appears to be able to identify TF functions and targets not easily accessible by either phylogenetic footprinting or ChIP-chip. We note that overexpression is only one type of artificial activation; other groups have fused TF DBDs to constitutive activation domains (14, 34). However, our results indicate that in many cases overexpression of the native protein, which may contain domains besides the DBD that are required for proper physiological function, will suffice for phenotypic activation.

The fact that the genes induced upon overexpression of TFs tend to include the bona fide targets argues that TF occupancy can be an important factor in the rate of transcription of many genes, because the simplest explanation is that overexpression increases occupancy by mass action. The observation that overexpression of TFs often causes growth inhibition suggests that cells are sensitive to aberrant activation of a variety of different pathways, and/or that there are signals that sense inappropriate pathway activation and reduce division rate. Consistent with this idea, our original study (18) also identified many signaling molecules that cause growth inhibition when overexpressed, presumably because they activate their targets similarly, in an

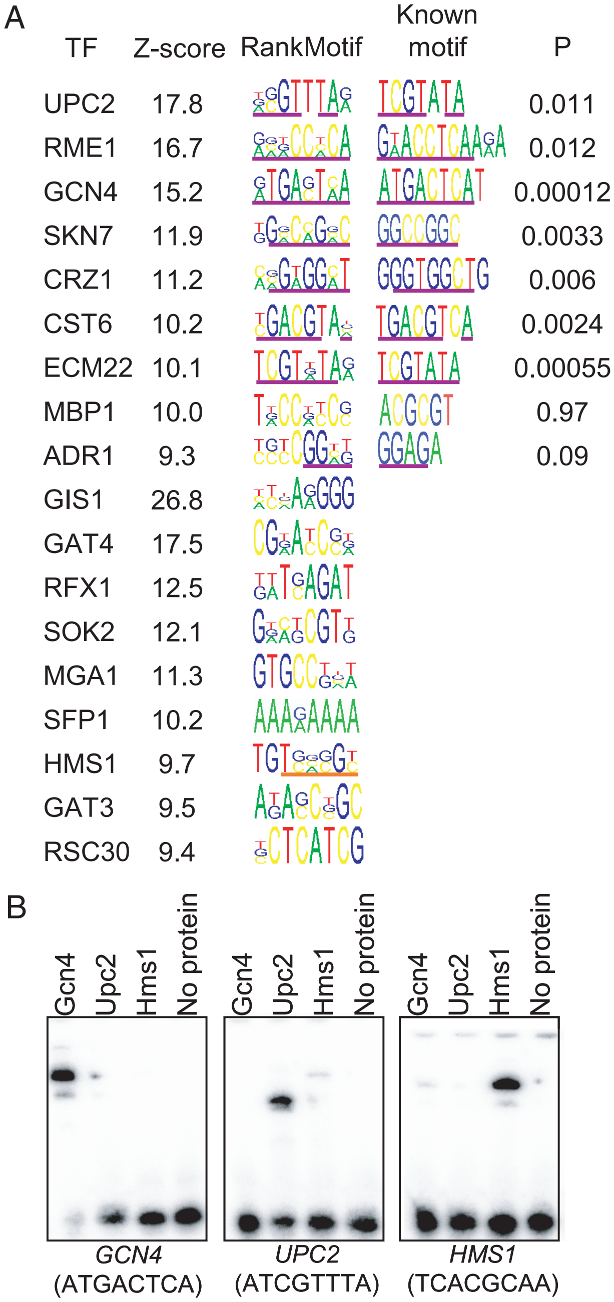

**Fig. 3.** Promoter analysis of differentially regulated genes in response to TF overexpression. (*A*) Motifs identified by RankMotif compared with known DNA-binding motifs for overexpressed TFs. Binding sites are displayed as logos in which the height of each letter is proportional to its weight in determining the motif. The purple underlined portion indicates bases consistent with the known binding site. The likelihood of the known motif matching the RankMotif consensus is given (formula and code are available on request). The orange underlined portion of the *HMS1* motif shows the six bases that match the gel-shifted segment in *B*. (*B*) Gel mobility-shift assays. The purified DBDs of Gcn4p, Upc2p, and Hms1p TFs were incubated with oligonucleotides containing two tandem copies of the motif sequence predicted by RankMotif. The same amount of purified MBP-TF DBD was used for each oligonucleotide in the binding reaction.

unregulated manner. Notably, when we generated microarray profiles of 23 well characterized TF overexpression strains that did not exhibit a reduced fitness when overexpressed a similar analysis to that shown here indicated that all were inactive (data not shown). These results indicate that overproduction of these TFs is not sufficient for their activation, although it remains possible that some of the TF-fusion proteins are nonfunctional. However, our simple initial phenotypic screen was sufficient to identify these constructs as unlikely to be worth pursuing.

Importantly, the general phenotypic activation approach described here, an initial screen for a visible phenotype upon TF overexpression, coupled to subsequent microarray analysis and a battery of statistical analyses, could be applied systematically in any organism for which an inducible transgene can be introduced, using commercially available custom oligonucleotide microarrays (35). There are already numerous instances in organisms ranging from microbes to mammals in which overexpression of TFs results in morphological abnormalities (36–38). It will be intriguing to determine whether expression profiling in these samples reveals induction of specific pathways whose genes contain binding sites for the TF in question. It will be equally fascinating to explore cases where pathways are also induced or repressed that do not appear to be direct targets of the TF, because such cases may result from transcriptional cascades. *HMS1*, for example, appears to positively regulate genes involved in several diverse pathways, including several that have dedicated TFs, and do not appear to contain Hms1p-binding sites in their promoters (Fig. 7*C*). In such situations, cause-and-effect relationships can often be determined by using epistasis analysis, a traditional genetic approach to mapping pathways that has itself been shown to be amenable to a microarray readout (17, 39).

## Methods

**Microarray Experiments.** Strains carrying 2$\mu$ plasmids that contain TFs regulated by the *GAL1* promoter were derived from a yeast overexpression array (18). For microarray experiments, the TF overexpression and empty vector control strains were grown concurrently in selective medium supplemented with 2% raffinose before induction with 2% galactose for 3 h, whereas TF deletion mutants and the isogenic WT strain were grown in synthetic medium supplemented with 2% dextrose. Procedures detailing culturing, RNA preparation, hybridizations, image acquisitions, and data processing for microarrays are described in Grigull *et al.* (22).

**Microarray Data Normalization.** Spatially detrended and Lowess-smoothed microarray data were obtained using protocols and microarrays as described (40). The output of this procedure is a normalized log ratio of intensity for each spot in the mutant strain versus the WT strain and the average log intensity for each spot in the two strains. The normalized log ratio itself is not a good measure of the significance of up- or down-regulation of the spot because the SD of the log ratios of unaffected spots decreases as a function of the average spot intensity. We transformed the log ratios into intensity-independent measures of significance of regulation, by calculating a *z*-score for each log ratio by dividing it by a robust estimate of the SD of unaffected spots (on the same array) with similar average intensities. Specifically, for each spot $i$ with a log ratio of $r_i$, its $z$-score, $z_i = (r_i - m_i)/s_i$, where $m_i$ and $s_i$ are the median and median absolute deviation, respectively, of the log ratio of all spots with average log intensities within 0.25 log units of spot $i$. These $z$-scores typically correspond to five times the $\log_2$(ratio). Microarray data before and after normalization and transformation will be available at the National Center for Biotechnology Information GEO database.

**WMW Tests for TF Target Enrichment.** Lists of yeast TF targets were downloaded from TRANSFAC (24). In total, binding data from Harbison *et al.* (4) and overexpression data from this study were available for 25 TFs in the TRANSFAC list. For each TF, we compared the log ratios of the TRANSFAC targets versus the nontargets in the overexpression assay with a two-sided WMW test. We also compared the Harbison *et al.* binding $P$ values of TRANSFAC targets versus nontargets by using a one-sided WMW test. For some TFs, Harbison *et al.* provide binding data for the TF under multiple growth conditions; in those cases, we assigned the TF the lowest $P$ value among all of the conditions and then multiplied the $P$ value by the number of conditions to correct for the multiple testing.

**WMW Tests for Gene Ontology Functional Enrichment.** Gene Ontology annotations (provided by the *Saccharomyces* Genome Database) were downloaded from www.geneontology.org on October 5, 2005. For each overexpression or deletion mutant, and for each Gene Ontology Biological Process (GO-BP) category containing >10 ORFs represented on our microarray, we used a two-sided WMW test to compare the $z$-scores of the ORFs annotated and unannotated in the given GO-BP category. We controlled for multiple testing by using the Benjamini–Hochberg procedure to calculate false discovery rate. In Fig. 2*C*, only the $P$ values of significantly enriched TF/GO-BP pairs are shown; any pair with a false discovery rate $> 0.01$ is assigned a $P$ value of 1 (i.e., appears as white).

**Extraction of Yeast Promoters.** Intergenic sequences were downloaded from the *Saccharomyces* Genome Database on October 13, 2005 (ftp://genome-ftp.stanford.edu/pub/yeast/sequence/genomic_sequence/intergenic). Promoters were defined as the intergenic sequence spanning the region immediately upstream of the start position of a given ORF to the end position of the upstream neighboring ORF. ORFs annotated as “dubious” were omitted from analysis. A FASTA file of promoter sequences is available on request.

**Motif Finding Using RankMotif.** RankMotif is a probabilistic inference algorithm that finds degenerate consensus sequences (taken as a motif model) that are overrepresented in the promoters of high-ranking ORFs in a ranked list. The input to RankMotif is a ranked list of ORFs and their associated promoter sequences. For a single TF, RankMotif searches for the highest-scoring degenerate consensus sequence. To model a stress response that is shared by multiple overexpression experiments, we introduced a shared motif model that is the same for all TFs. By incorporating the shared motif model, the score of an individual model is the maximum of the original score and the score calculated based on the sum of the ranks of all ORFs whose promoters contain either or both motifs. RankMotif iterates between updating the shared motif model given the current individual motif models (by modifying positions and shifting the alignment of the motif right and left by a single base), and updating the individual motif models given the current shared motif model. The search ends when the current state has a higher score than all possible updates. In the experiments described here, we also attempted to avoid some of the drawbacks of greedy search by also maintaining and updating a set of 19 suboptimal motif models for each TF and for the shared motif model. To allow for strand preference, we also ran RankMotif on three different sets of promoters for the ORFs consisting of the sense strands, antisense strands, and both. RankMotif was run for five iterations for each of these three promoter sets. We found the top specific and nonspecific motifs and scores for the three strand options. The individual motif models reported were those that had the highest score among the RankMotif output for the three promoter sets. Full technical details will be described elsewhere (Q.D.M., unpublished work).

**Purification of DBD and Gel Mobility Shifts.** The DBDs and 10–15 flanking amino acids of Gcn4p (amino acids 206–281), Upc2p (amino acids 1–120), and Hms1p (amino acids 256–360) were PCR-amplified and fused at their N termini to the maltose-binding protein (MBP) by cloning into the pMAL-C2 vector. The fusion proteins were expressed in BL21 (DE3) cells and purified with amylose resin (NEB, Beverly, MA). The gel-shift probes consisted of two tandem copies of the 8-mer motif representing the TF binding site followed by 16 nt of nonyeast sequence common to all of the probes. Sequences were as follows: *GCN4*, 5′-*ATGACTCAATGACTCA*CCTCGGCTGCAGGTAC-3′; *UPC2*, 5′-*ATCGTTTAATCGTTTA*CCTCGGCTGCAGGTAC-3′; and *HMS1*, 5′-*TCACGCAATCACGCAA*CCTCGGCTGCAGGTAC-3′. For the binding reaction, 0.1 pmol of 5′-$^{32}$P-end-labeled probe and purified MBP-TF DBD was incubated with gel-shift reaction buffer (10 mM Hepes, pH 7.8/75 mM KCl/2.5 mM $MgCl_2$/1 mM DTT/3% Ficoll) at room temperature in a 10-μl binding reaction. Final protein concentrations were: Gcn4p-DBD, 119 nM; Upc2p-DBD, 107 nM; and Hms1p-DBD, 129 nM. After 1 h, 3 μl of 20% Ficoll (Sigma, St. Louis, MO) was added, and the reaction was loaded onto a 5% nondenaturing acrylamide gel and then visualized with a PhosphorImager (Bio-Rad, Hercules, CA). The same amount of purified MBP-TF DBD was used for each probe in the binding reaction.

**Data Availability.** All microarray data (before and after *z*-score transformation), spreadsheets underlying the figures, lists of known TF targets, WMW scores for all functional categories in all experiments, a table of properties of the TFs, and algorithms for computing the significance of motif matches in Fig. 3*A* are available on request. Microarray data will be available at the National Center for Biotechnology Information GEO database.

This work was supported by grants from the Natural Sciences and Engineering Research Council (to T.R.H., C.B., and B.J.F.) and funds from Genome Canada through the Ontario Genomics Institute (to T.R.H., C.B., and B.J.A.). G.C. was supported by a Charles H. Best Postdoctoral Fellowship. Q.D.M. was supported by a Natural Sciences and Engineering Research Council postdoctoral fellowship. R.S. was supported by a National Cancer Institute of Canada Terry Fox Foundation research studentship.

1. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. (2003) *Nature* **423,** 241–254.
2. Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A. & Johnston, M. (2003) *Science* **301,** 71–76.
3. Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., *et al.* (2005) *Genome Res.* **15,** 1034–1050.
4. Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., *et al.* (2004) *Nature* **431,** 99–104.
5. Gray, P. A., Fu, H., Luo, P., Zhao, Q., Yu, J., Ferrari, A., Tenzen, T., Yuk, D. I., Tsung, E. F., Cai, Z., *et al.* (2004) *Science* **306,** 2255–2257.
6. Chua, G., Robinson, M. D., Morris, Q. & Hughes, T. R. (2004) *Curr. Opin. Microbiol.* **7,** 638–646.
7. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. (1998) *Nat. Biotechnol.* **16,** 939–945.
8. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) *Nat. Genet.* **22,** 281–285.
9. Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A. & Bulyk, M. L. (2004) *Nat. Genet.* **36,** 1331–1339.
10. Liu, X., Noll, D. M., Lieb, J. D. & Clarke, N. D. (2005) *Genome Res.* **15,** 421–427.
11. Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al.* (2000) *Science* **290,** 2306–2309.
12. Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M. & Brown, P. O. (2001) *Nature* **409,** 533–538.
13. Gao, F., Foat, B. C. & Bussemaker, H. J. (2004) *BMC Bioinformatics* **5,** 31.
14. Devaux, F., Marc, P., Bouchoux, C., Delaveau, T., Hikkel, I., Potier, M. C. & Jacq, C. (2001) *EMBO Rep.* **2,** 493–498.
15. DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997) *Science* **278,** 680–686.
16. Madhani, H. D., Galitski, T., Lander, E. S. & Fink, G. R. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 12530–12535.
17. Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y. D., Dai, H., Walker, W. L., Hughes, T. R., *et al.* (2000) *Science* **287,** 873–880.
18. Sopko, R., Huang, D., Preston, N., Chua, G., Papp, B., Kafadar, K., Snyder, M., Oliver, S. G., Cyert, M., Hughes, T. R., *et al.* (2006) *Mol. Cell* **21,** 319–330.
19. Lorenz, M. C. & Heitman, J. (1998) *Genetics* **150,** 1443–1457.
20. Natarajan, K., Meyer, M. R., Jackson, B. M., Slade, D., Roberts, C., Hinnebusch, A. G. & Marton, M. J. (2001) *Mol. Cell. Biol.* **21,** 4347–4368.
21. Hinnebusch, A. G. (2005) *Annu. Rev. Microbiol.* **59,** 407–450.
22. Grigull, J., Mnaimneh, S., Pootoolal, J., Robinson, M. D. & Hughes, T. R. (2004) *Mol. Cell. Biol.* **24,** 5534–5547.
23. Vik, A. & Rine, J. (2001) *Mol. Cell. Biol.* **21,** 6395–6405.
24. Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., *et al.* (2003) *Nucleic Acids Res.* **31,** 374–378.
25. Liu, X., Brutlag, D. L. & Liu, J. S. (2001) *Pac. Symp. Biocomput.* **6,** 127–138.
26. Roberts, R. L. & Fink, G. R. (1994) *Genes Dev.* **8,** 2974–2985.
27. Lorenz, M. C., Cutler, N. S. & Heitman, J. (2000) *Mol. Biol. Cell* **11,** 183–199.
28. de Montigny, J., Kern, L., Hubert, J. C. & Lacroute, F. (1990) *Curr. Genet* **17,** 105–111.
29. Mao, C., Xu, R., Bielawska, A. & Obeid, L. M. (2000) *J. Biol. Chem.* **275,** 6876–6884.
30. Kirsch, D. R. & Whitney, R. R. (1991) *Infect. Immun.* **59,** 3297–3300.
31. Varma, A., Edman, J. C. & Kwon-Chung, K. J. (1992) *Infect. Immun.* **60,** 1101–1108.
32. Goldstein, A. L. & McCusker, J. H. (2001) *Genetics* **159,** 499–513.
33. Prasad, T., Saini, P., Gaur, N. A., Vishwakarma, R. A., Khan, L. A., Haq, Q. M. & Prasad, R. (2005) *Antimicrob. Agents Chemother.* **49,** 3442–3452.
34. Webster, N., Jin, J. R., Green, S., Hollis, M. & Chambon, P. (1988) *Cell* **52,** 169–178.
35. Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., *et al.* (2001) *Nat. Biotechnol.* **19,** 342–347.
36. Duncan, M. K., Xie, L., David, L. L., Robinson, M. L., Taube, J. R., Cui, W. & Reneker, L. W. (2004) *Invest. Ophthalmol. Visual Sci.* **45,** 3589–3598.
37. Seufert, D. W., Prescott, N. L. & El-Hodiri, H. M. (2005) *Dev. Dyn.* **232,** 313–324.
38. Hochedlinger, K., Yamada, Y., Beard, C. & Jaenisch, R. (2005) *Cell* **121,** 465–477.
39. Van Driessche, N., Demsar, J., Booth, E. O., Hill, P., Juvan, P., Zupan, B., Kuspa, A. & Shaulsky, G. (2005) *Nat. Genet.* **37,** 471–477.
40. Mnaimneh, S., Davierwala, A. P., Haynes, J., Moffat, J., Peng, W. T., Zhang, W., Yang, X., Pootoolal, J., Chua, G., Lopez, A., *et al.* (2004) *Cell* **118,** 31–44.