Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation

Frederick P. Roth[1–3†], Jason D. Hughes[1,2†], Preston W. Estep[2], and George M. Church[1,2*]

*[1]Harvard University Graduate Biophysics Program and [2]Harvard Medical School Department of Genetics, Boston, MA 02115. [3]Current address: Millennium Information, Cambridge, MA 02139. *Corresponding author (e-mail: church@salt2.med.harvard.edu). †These authors contributed equally to this work.*

Received 2 April 1998; accepted 10 August 1998

Whole-genome mRNA quantitation can be used to identify the genes that are most responsive to environmental or genotypic change. By searching for mutually similar DNA elements among the upstream noncoding DNA sequences of these genes, we can identify candidate regulatory motifs and corresponding candidate sets of coregulated genes. We have tested this strategy by applying it to three extensively studied regulatory systems in the yeast *Saccharomyces cerevisiae*: galactose response, heat shock, and mating type. Galactose-response data yielded the known binding site of Gal4, and six of nine genes known to be induced by galactose. Heat shock data yielded the cell-cycle activation motif, which is known to mediate cell-cycle dependent activation, and a set of genes coding for all four nucleosomal proteins. Mating type α and a data yielded all of the four relevant DNA motifs and most of the known a- and α-specific genes.

Keywords: functional genomics, gene expression

Complete DNA sequence is now known for more than 10 different organisms[1]. For even the most intensely studied of these organisms, a large fraction of genes is completely uncharacterized—about 40% and 50% for *Escherichia coli* and *Saccharomyces cerevisiae*, respectively[2,3]. Furthermore, annotation of noncoding regions has typically lagged behind discovery and prediction of gene function. Given that sequence elements in noncoding regions often control gene expression, and that knowing a gene's place in the larger regulatory network of a cell is essential to understanding its function, it is critical that we develop methods for rapidly characterizing noncoding regions.

A common approach to the discovery of regulatory elements entails the construction of a series of deletions or replacements in the upstream intergenic region of a gene, followed by an assay for altered regulation. An efficient method for predicting the most likely locations of regulatory sequences could guide these experiments more quickly to the sought-after elements. Given a set of genes "enriched" for coregulated members (obtained, for example, by genetic evidence), a search for conserved upstream sequence elements can predict the location of gene regulatory sequences[2].

Recently, it has become possible to measure the abundance of mRNA transcripts on a whole-genome scale[5–9]. By comparing transcript levels between different conditions or different strains we can find the set of genes whose transcript levels respond to a difference in environment or genotype. With this set of genes in hand, a number of questions naturally arise: Which of these changes in expression constitutes a primary response to an environmental change and which are indirect effects? Which are most critical for adaptation to a new condition? By what mechanisms are changes in transcript abundance achieved? What DNA or RNA sequence elements mediate the regulation of transcript abundance? It is the last question that we seek to address here.

Given a set of induced (or repressed) genes, one can search the regions upstream of translation start for short DNA sequence motifs, i.e., aligned sets of short, conserved DNA elements that are candidate DNA regulatory motifs. This search for regulatory motifs does not depend on prior information about gene regulatory mechanism and can be contrasted with approaches that search in upstream regions of induced (or repressed) genes for new examples of known DNA regulatory motifs[8].

Although they have been less extensively studied, conserved sequence elements in a gene's upstream region may also be determinants of mRNA stability[10,11] or even sites for regulation by antisense transcripts[12]. Regardless of mechanism, a highly conserved DNA sequence upstream of genes with similar expression responses is of interest. Similarly expressed genes that share a conserved upstream DNA motif constitute a candidate set of coregulated genes.

To test our strategy of combined expression analysis and upstream sequence alignment, we examined three extensively studied, transcriptionally regulated systems in the yeast *S. cerevisiae*: galactose utilization, heat shock response, and mating type regulation. To examine these three systems, we measured mRNA transcript abundance in *S. cerevisiae* on a whole-genome scale in each of four different cultures, which allowed three comparisons to be made: (1) growth on galactose vs. glucose, (2) strains of mating type α vs. mating type a, and (3) continuous growth at 30°C vs. 30°C growth followed by a 39°C heat shock. Expression was measured for each of these comparisons using photolithographically synthesized oligonucleotide microarrays ("chips")[7,9]. Change in transcript abundance for each open reading frame (ORF) was calculated for each comparison (Fig. 1).

Results and discussion

Examining upstream noncoding DNA sequence. We examined sets of upstream DNA sequences corresponding to (1) the top 10 ORFs as ranked by ratio of first-condition to second-condition abundance (e.g., ratio of galactose to glucose expression), (2) the top 10 ORFs as ranked by ratio of second-condition to first-condition abundance (e.g., ratio of glucose to galactose expression), and (3) the combination of the two preceding ORF sets. The rationale for examining the combined set of upstream regions is that a single regulatory motif

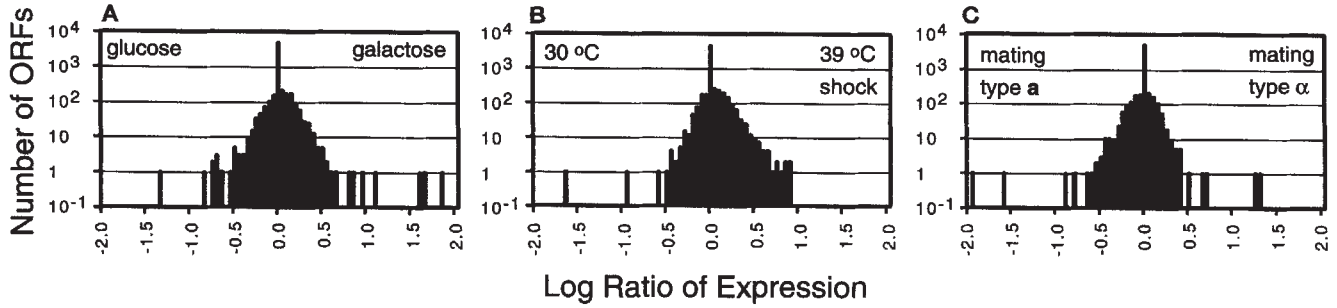

Figure 1. Histograms of change in expression level for each of three whole-genome expression comparisons. (A) Growth in galactose vs. glucose. (B) Growth after heat shock vs. 30°C. (C) Growth of mating type α vs. a. Transcripts with undetectable abundance in both conditions were assigned a log ratio value of 1.

may act as either a negative or a positive regulator depending on its sequence context. The upstream DNA sequence corresponding to each ORF was bounded at the 3′ (or downstream) end by the ORF's translation start. The 5′ end of the upstream region was bounded by the translation start or stop of the nearest upstream ORF, except that this boundary was never >600 DNA bp or <300 bp from translation start. The choice of upstream region boundaries is justified by an examination of those *S. cerevisiae* regulatory sites listed in the TRANSFAC database for which locations are given relative to translation start. Eighty-five percent (94 of 110) of these sites lie between 0 and 600 bases upstream of translation start.

Several algorithms for discovering recurring motifs in unaligned sequences have been developed. Those that are capable of automatically producing alignments containing multiple sites from a single input sequence include Gibbs Motif Sampling (GMS), CoreSearch, three CONSENSUS variants[13], and MEME[14]. While each method has its advantages and limitations with respect to our intended application, we chose GMS[15,16] to serve as our starting point for further development as we believed it to have the most flexible and exhaustive search methodology. There are several major distinctions between the application used here (called "AlignACE," for *Align*s *N*ucleic *A*cid *C*onserved *E*lements) and GMS, as implemented by Neuwald et al.[15] AlignACE has been optimized for finding multiple motifs (via an iterative masking procedure) and for alignment of DNA sequences (by automatic consideration of both strands). It also scores alignments by frequency of occurrence in the intergenic DNA sequence of a given genome.

AlignACE was applied to each of the sets of upstream DNA sequence described above. Of the many resulting DNA site alignments, we considered only those motifs that: (1) exceeded a threshold alignment score (a measure of "goodness" of sequence alignment), and (2) had an occurrence score (a measure of the fraction of ORFs in the *S. cerevisiae* genome with matching upstream sites) below 1%. The latter criterion requires that motifs be selective—that is, occur infrequently among upstream regions.

Those motifs that passed both alignment and occurrence score criteria were then compared with motifs one might have expected to find, with varying levels of confidence, given the relevant literature. For this purpose, we developed an objective measure of similarity between sequence motifs.

Galactose vs. glucose comparison. Using the set of upstream regions corresponding to the 10 ORFs with transcripts ranked most increased in galactose relative to glucose, we identified a motif, which we called gal-1, that matched the galactose upstream activation sequence (UAS_G) motif. UAS_G is known to regulate galactose-utilization genes via the Gal4/Gal80 activation complex[17]. No motif passing both alignment and occurrence criteria was obtained when upstream regions corresponding to the 10 ORFs with transcripts ranked most increased in glucose relative to galactose were used. Another UAS_G-like motif, gal-glu-1, was obtained when the preceding two upstream region sets (corresponding to a total of 20 ORFs) were combined (Table 1 and Fig. 2).

Table 1. DNA motifs found and expected.

Comparison	Found by AlignACE	Alignment score	% occurence	Similar motifs
galactose vs. glucose	gal-1	33.1	0.16	UAS_G
	gal-glu-1	24.9	0.20	UAS_G
heat shock vs. 30°C	39C-1	5.1	0.04	
	30C-1	40.1	0.26	CCA
	30C-2	5.5	0.44	-
	39C-30C-1	30.1	0.20	CCA
	39C-30C-2	8.5	0.10	
mating type α vs. mating type a	mtα-1	8.9	0.22	P Box
	mta-1	8.5	0.10	-
	mta-2	5.0	0.20	-
	mta-3	28.1	0.62	α2-binding
	mtα-mta-1	20.7	0.68	α2-binding
	mtα-mta-2	5.3	0.26	PRE
	mtα-mta-3	8.6	0.54	-
	mtα-mta-4	5.3	0.62	Q Box

Comparison	Expected motif	DNABP	Reference
galactose vs. glucose	UAS_G	Gal4p/Gal80p	40
	URS_G	Mig1p	40
	Rap1p-binding	Rap1p	40
	Gcr1p-binding	Gcr1p	40
heat shock vs. 30°C	HSE	HSF	42
	STRE	Msn2p/Msn4p	43, 44
	CCA	?/Hir1p/Hir2p	19
	NEG	?	20
	MCB	Mbp1p	21
	SCB	Swi4p/Swi6p	21
	ECB	Mcm1p	22
mating type α vs. mating type a	P Box	Mcm1p	23
	Q Box	Matα1p	23
	α2-binding	Matα2p	23
	PRE	Ste12p	23

Similar motifs: those expected motifs found to be similar by objective criteria; DNABP: the protein that binds an element where known.

Heat shock. Upstream regions of the 10 ORFs whose transcript levels increased the most in the heat-shocked culture relative to the 30°C culture yielded a single motif, 39C-1. Upstream regions of the 10 ORFs with transcripts ranked most increased in the 30°C relative to the heat-shocked culture yielded two motifs. The first of these, 30C-1, matched the cell cycle activation (CCA) motif, a known activator of histone genes[19,20]. The second motif, 30C-2, has not been described. When the combined set of upstream regions was used, another CCA-like motif—39C-30C-1—was identified along with 39C-30C-2,which was similar to 39C-1.

Mating type. When upstream regions of the 10 ORFs whose transcript levels increased the most in mating type α relative to

type a were examined, the motif mtα-1 was found. mtα-1 matched the P Box and the early cell-cycle box (ECB), as well as the Gcr1p-binding site and the heat shock element (HSE). ECB mediates M/G1-specific activation[21,22] and the P Box regulates mating type-specific genes[23]. The P Box and ECB both bind Mcm1p[23]. Using upstream regions of the 10 ORFs whose transcript levels increased the most in mating type a, relative to type α, three motifs emerged. The first two of these, mta-1 and mta-2 have not been described. The third motif, mta-3, matched the binding site of Matα2p[21].

When upstream regions from the combined set of 20 ORFs with most altered transcript abundance between mating type α and a were examined, four motifs emerged. The first of these, mtα-mta-1, matched the known Matα2p-binding site. The second, mtα-mta-2, matched the pheromone-response element (PRE)—the known binding site of Ste12p[23], an activator of mating type-specific genes. A weak second match between mtα-mta-2 to the PRE motif suggested a conserved spacing of 7 bp between PRE elements. The third motif, mtα-mta-3, bore some resemblance to the PRE motif, but had a similarity score below the threshold applied. The fourth motif, mtα-mta-4, corresponded to the Q Box element, which binds Matα1p and mediates activation of α-specific genes[23]. Motifs similar to the PRE and Q Box motifs were not found by examining only upstream regions of the combined set of 20 ORFs or either set of 10 ORFs. The PRE element confers inducibility by either α or a mating pheromone, and was previously known to be present upstream of both mating type α- and a-specific genes. The Q Box, on the other hand, was expected only upstream of α-specific genes so that finding a matching site upstream of *STE2* was unexpected. A closer examination of this *STE2* site revealed a T instead of the highly conserved A at the eighth position of the Q Box motif, so that the match was not a strong one.

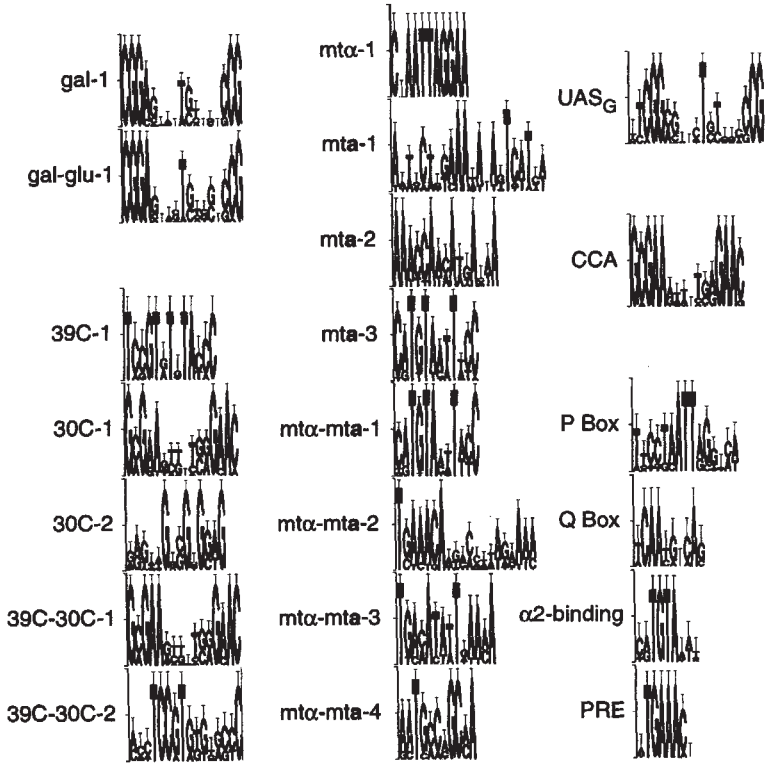

Figure 2. Sequence logos for DNA motifs found using AlignACE (first and second columns) and similar motifs that might have been expected a priori (third column). The height of each letter is proportional to its frequency, and the letters are sorted so the most common one is on top. The height of the entire stack signifies the information content of the sequences at that position, with information content at each position varying between 0 and 2 bits[18].

Comparison with other studies. Motifs common to many genes—for example, the TATA box—were neither found nor expected as these are excluded by the occurrence score constraint discussed above. In the case of transcripts less abundant in galactose than glucose, motifs corresponding to Rap1p- or Gcr1p-binding sites might be expected, but neither of these was found. Rap1p and Gcr1p are general transcription factors with diverse roles, including regulation of glycolytic enzymes and ribosomal proteins[24]. In the case of transcripts more abundant in galactose than glucose, we expected to find URS$_G$ (bound by Mig1p[24]), but did not. As expected, our procedure did find UAS$_G$, an essential regulatory element for galactose-utilization genes.

The HSE and stress response promoter elements (STREs), known to mediate heat shock response[25,26], were notably absent from the motifs found by AlignACE. Heat shock is known to have broad effects, including transient cell-cycle arrest in G1[27]. As a result, we might have expected to find genes with cell cycle–specific expression among genes affected by heat shock. Histone genes are strongly transcribed during S phase, and are regulated both by a negative regulatory sequence (NEG) and activated by the CCA motif[19,20]. Not found among heat shock data were the NEG motif, the Swi4/6p-dependent cell cycle box (SCB), the MluI cell cycle box (MCB) motifs (which regulate G1/S-specific transcription), or the ECB element[21,22]. AlignACE did find the CCA motif among the set of ORFs with transcripts decreased in heat shock, a set which contained several histone genes.

The α2 operator, P Box, PRE, and Q Box elements represent the complete set of DNA elements responsible for regulation of mating type-specific genes[23]. All four of these elements were found by AlignACE.

It is likely that the six motifs found by AlignACE that we did not identify with previously known motifs are false positives. Six motifs represent no more than the average number of false positives one might have expected, as determined below.

False-positive and false-negative motifs. The alignment and occurrence score criteria used here, 5 and 1%, respectively, were chosen permissively, so that few biologically relevant motifs would be excluded. To ensure that the alignment criterion was sufficiently permissive, we searched for conserved motifs among upstream sequences corresponding to 1000 randomly chosen sets of 10 ORFs. A maximum of three motifs was sought from each intergenic sequence set. Each randomly chosen sequence set returned at least one motif with a score greater than our threshold of five, and in 674 cases, all three motifs returned had scores greater than the permissive. An occurrence score of 1% seemed to be a permissive threshold, given our desire for motifs that do not match too frequently in upstream regions.

Because different users of this method will have varying tolerance for false-positive motifs, it is advantageous to know how the expected number of false-positive motifs depends on chosen alignment and occurrence score thresholds. We estimated the number of expected false positives by randomly generating 100 sets of 10 ORFs, obtaining the upstream DNA sequence corresponding to these ORFs, and plotting the expected number of false-positive motifs as a function of alignment and occurrence score thresholds (Fig. 3). For the permissive thresholds, an average of 1.7 motifs (number of false positives is Poisson-distributed with a coefficient of dispersion of 0.9) was obtained from randomly chosen upstream sequence sets. Eight motifs identified from 6 sets of 10 ORFs meet the permissive thresholds. Four of these match known regulatory motifs. The chance that a motif is a false positive decreases with increasing

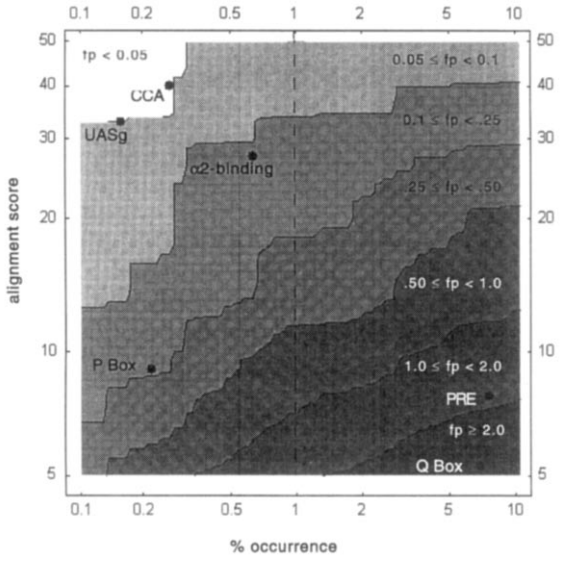

Figure 3. Expected number of false-positive motifs (fp) as a function of alignment and occurrence scores, based on analysis of upstream sequences corresponding to randomly chosen sets of ten ORFs. True-positive motifs found upstream of one of the top 10 sets of increased or decreased transcripts are overlaid. Only those motifs found to be similar to motifs that might have been expected a priori are shown. The dashed line corresponds to the permissive occurrence score criterion discussed in the text.

Table 2. Change in transcript abundance for galactose versus glucose.

ORF ID	Gene	Change	gal-1	gal-glu-1
YBR020W	**GAL1**	**>1.8**	**5**	**4**
YBR018C	**GAL7**	**>1.6**	**2**	**2**
YBR019C	**GAL10**	**>1.6**	**5**	**4**
YOR120W	**GCY1**	**>1.1**	**1**	**1**
YLR081W	**GAL2**	**>0.9**	**4**	**4**
YPL066W	**-**	**>0.8**	**1**	**1**
YPL067C	**-**	**>0.8**	**1**	**1**
YMR318C	**-**	**0.6**	**1**	**1**
YNL015W	**PBI2**	**>0.6**	**-**	**-**
YOL058W	**ARG1**	**0.5**	**-**	**-**
YDR009W	GAL3	>0.5	2	2
YML051W	*GAL80*	*ND*	*2*	*2*
YMR105C	*PGM2*	*ND*	*1*	*1*
YBR184W	*MEL1*	*ND*	-	-
YER178W	**PDA1**	**-0.5**	**-**	**-**
YBR011C	**IPP1**	**-0.6**	**-**	**-**
YER190W	**-**	**-0.7**	**-**	**-**
YBR106W	**PHO88**	**-0.7**	**-**	**-**
YFL045C	**SEC53**	**-0.7**	**-**	**-**
YGL030W	**RPL30**	**-0.7**	**-**	**-**
YOL154W	**-**	**-0.8**	**-**	**-**
YHR094C	**HXT1**	**-0.8**	**-**	**-**
YGL189C	**RPS26A**	**-0.9**	**-**	**-**
YDR345C	**HXT3**	**-1.4**	**-**	**-**

The top 10 increased or decreased transcripts are in bold type; transcripts not among the top 10 but that had greater than 0.3 absolute value of change and an upstream match to either gal-1 or gal-glu-1 are in plain text. Transcripts regulated by the Gal4p/Gal80p complex, but that had less than 0.3 absolute change, are in italics. Positive change indicates higher abundance in galactose than glucose.

alignment score so that if a threshold score of 20 is applied, the mean number of false positives is 0.2. Three motifs identified from 6 sets of 10 ORFs meet this most stringent threshold. All three of these match known regulatory motifs. These methods can then be applied—at the discretion of the user—to either generate an inclusive set of testable hypotheses with a significant false-positive rate, or to more confidently predict a subset of biologically relevant motifs. A sensible alternative to using fixed score thresholds is the choice of a threshold based on a desired number of false positives, i.e., some combination of alignment and occurrence score corresponding to a contour in Figure 3.

We were interested in understanding whether and why our approach had false negatives, i.e., why it did not find all of the motifs we might have expected from the literature. There are several possible reasons why a given motif might not be found: (1) The motif is not, in fact, involved in differential regulation in the examined cultures, i.e., we should not have expected to find the motif; (2) the motif does regulate transcription between examined cultures, but there are few or no sites found among the upstream regions of the top 10 differentially regulated transcripts; or (3) multiple sites are present upstream of the top 10 differentially expressed genes, but were not aligned by AlignACE or scored below the permissive criteria. We can address the third reason by searching upstream of the top 10 differentially regulated genes for sites matching the motif (as defined in the literature). Among upstream regions corresponding to the 10 transcripts most increased in galactose, the URS motif matched only one site (between *GAL1* and *GAL10*). Among upstream regions corresponding to the 10 transcripts most decreased in galactose, the Rap1-binding motif matched four sites (all upstream of *RPL30*) but the Gcr1-binding motif did not match any site. Among the 10 genes most increased in heat shock, HSE did not match any sites, but STRE matched 11 sites (five, two, three, and one site upstream of *HSP12, YDR453C, HSP26,* and *YDR533C,* respectively). Among the 10 genes most decreased in heat shock, the MCB motif matched two sites (between *HTA2* and *HTB2*), the NEG motif matched one site (between *HHT1* and *HHF1*), and neither SCB nor ECB matched any site. We conclude from this analysis that STRE is the only false-negative motif that can clearly be attributed to a shortcoming of AlignACE.

Candidate sets of coregulated genes. Once a candidate regulatory motif has been found, we are particularly interested in those ORFs with both altered transcript abundance and also an associated upstream match to the motif. First, we search the complete *S. cerevisiae* genome to find ORFs with upstream matches to each motif in the 600 bp region upstream of translation start. Second, we consider the intersection of this set of ORFs with those ORFs with transcripts that changed more than twofold in abundance between conditions. We consider ORFs with both of these properties to be good candidates for membership in a transcriptionally coregulated set of genes.

ORFs whose transcript abundance changes more than twofold between galactose and glucose, and have an upstream match to either of the UAS_G-like motifs gal-1 and gal-glu-1 are shown in Table 2, along with the remaining genes known previously to be regulated by UAS_G. There are nine genes known previously to be regulated by the Gal4p/Gal80p complex: *GAL1, GAL2, GAL3, GAL7, GAL10, GAL80, GCY1, MEL1,* and *PGM2.* The candidate set derived from Table 2 contains six of these nine. One of the missing three is *MEL1,* which codes for alpha-galactosidase. The strain used for this experiment (FY4) is a *mel*- strain. Furthermore, we could not find a good match to UAS_G in the *MEL1* upstream sequence from this strain, so that we should not necessarily have expected to find *MEL1* in our candidate set. Also missing from our candidate set were *PGM2* and *GAL80,* which are known to be Gal4p/Gal80p regulated. For both of these genes, absolute expression levels were too low in both galactose and glucose for an accurate estimate of change in expression.

The candidate set for galactose vs. glucose contains three ORFs not previously thought to be regulated by Gal4p/Gal80p. Two of these, *YPL066W* and *YPL067C,* are particularly interesting as they have approximately the same measured change in expression and are divergently transcribed from the same intergenic region that contains a single match to each of the UAS_G-like motifs. Another site in this intergenic region matches the consensus CGG(N)^{10}CCG, to which weak Gal4-binding has also been shown in vitro[28]. Both *YPL066W* and *YPL067C* are of unknown function[29], with no signif-

Table 3. Change in transcript abundance for heat shock vs. 30°C.

ORF ID	Gene	Change	39C-1	30C-1	30C-2	39C-30C-1	39C-30C-2
YFL014W	**HSP12**	**0.9**	-	-	-	-	-
YDR453C	**-**	**>0.9**	-	-	-	-	-
YPL223C	**GRE1**	**>0.8**	1	-	-	-	-
YGL121C	**-**	**>0.8**	-	-	-	-	-
YLR303W	**MET17**	**0.8**	-	-	-	-	-
YBR072W	**HSP26**	**0.8**	2	-	-	-	2
YGR256W	**GND2**	**>0.8**	1	-	-	-	1
YLR178C	**TFS1**	**0.7**	-	-	-	-	-
YDR533C	**-**	**0.7**	-	-	-	-	-
YDR019C	**GCV1**	**0.6**	1	-	-	-	1
YLL039C	UBI4	0.6	2	-	-	-	1
YDR070C	-	0.5	-	2	-	-	-
YER062C	HOR2	0.5	-	1	-	-	-
YER103W	SSA4	0.5	1	-	-	-	-
YHR007C	ERG11	0.5	1	-	-	-	1
YER042W	-	0.4	1	-	-	-	-
YDL223C	-	0.4	-	-	1	-	-
YGR038W	ORM1	0.4	-	-	1	-	-
YDL048C	STP3	>0.3	1	-	-	-	-
YMR090W	-	>0.3	1	-	-	-	-
YGR248W	SOL4	0.3	1	-	-	-	1
YOR185C	GSP2	0.3	1	-	-	-	-
YOR259C	CRL13	0.3	-	-	1	-	-
YNL241C	ZWF1	0.3	-	-	1	-	-
YNL156C	-	0.3	1	-	-	-	-
YPL135W	-	0.3	1	-	-	-	-
YHR057C	CYP2	0.3	1	-	-	-	-
YNL031C	*HHT2*	*ND*	-	*2*	*1*	*2*	*1*
YDR224C	*HTB1*	*-0.1*	-	*4*	*1*	*2*	-
YDR225W	*HTA1*	*-0.2*	-	*4*	*1*	*2*	-
YBL015W	ACH1	-0.3	-	-	-	-	1
YNL030W	HHF2	-0.4	-	2	1	2	1
YIR034C	**LYS1**	**-0.4**	-	-	-	-	-
YBR009C	**HHF1**	**<-0.4**	-	4	3	4	-
YFR015C	**GSY1**	**<-0.5**	2	1	1	-	-
YBL003C	**HTA2**	**-0.5**	-	5	-	4	-
YBL002W	**HTB2**	**-0.5**	-	5	-	4	-
YBL072C	**RPS8A**	**-0.5**	-	1	-	-	1
YBR010W	**HHT1**	**-0.5**	-	4	3	4	-
YGR234W	**YHB1**	**<-0.6**	2	-	2	-	1
YOL154W	**-**	**-1.0**	-	-	-	-	-
YJL052W	**TDH1**	**<-1.7**	-	-	1	-	-

The top 10 increased or decreased transcripts are in bold type; transcripts not among the top 10 but that had greater than 0.3 absolute change and an upstream match to one of the heat shock vs. 30°C-related motifs found by AlignACE are in plain text. Transcripts that code for nucleosomal proteins, but that had an absolute change of less than 0.3 are in italics. Positive change indicates higher transcript abundance in heat shock than 30°C.

icant homology to any gene of known function. Also in the set was *YMR318C*, which has a strong homology to zinc-containing alcohol dehydrogenases[29]. We are currently exploring the possibility that these ORFs are Gal4p/Gal80p-regulated genes.

ORFs that have a measured transcript abundance change of more than twofold between heat shock and 30°C, and have an upstream match to at least one of the motifs found through this comparison are shown in Table 3, along with the remaining histone genes. There are eight genes—four nearly identical pairs—in *S. cerevisiae* that code for the four nucleosomal proteins: *HTA1, HTA2, HTB1, HTB2, HHT1, HHT2, HHF1*, and *HHF2*. The candidate set of coregulated genes with matches to both of the CCA-like motifs (30C-1 and 39C-30C-1) contains five of these genes—collectively coding for the complete set of *S. cerevisiae* nucleosomal proteins. Our results alone do not necessarily indicate a role for the CCA motif in heat shock regulation or in cell cycle regulation, as we seek any motif common to a set of upstream regions without information about which motif (if any) might be responsible for the observed changes in expression. *RPS8A* (coding for ribosomal protein rp19), *HOR2* (DL-glycerol-3-phosphatase), *YDR070C* (an ORF of unknown function), and GSY1 (glycogen synthase) have upstream matches to one of the CCA-like motifs (30C-1)[29]. None of these genes have previously been noted to be heat shock repressed or to be regulated by CCA. However, *GSY2*—a homolog of *GSY1*—is reportedly induced by heat shock[30]. This suggests, together with our data showing reduced abundance of GSY1 in heat shock, that *GSY2* is a thermotolerant variant of glycogen synthase. Along with the other histone genes, we found the gene *HHO1* (histone H1) to be reduced in abundance by more than twofold in heat shock. *HHO1* does not, however, contain an upstream match to the CCA motif.

Another candidate gene set can be derived from heat shock data, using those ORFs in Table 2B with upstream matches to either of the two mutually similar motifs 39C-1 and 39C-30C-2. This set contains the genes *HSP26, GND2, GCV1, UBI4, ERG11, SOL4*, and *YHB1*. Of these, only *HSP26* and *UBI4* have been reported to be differentially expressed as a result of heat shock[25]. We can see no apparent rationale for coregulation of this set of genes or of the set derived from 30C-2. None of the motifs 39C-1, 30C-2, or 39C-30C-2 have been implicated as regulatory elements.

ORFs that have a measured transcript abundance change of more than twofold between mating types α and a, and have an upstream match to at least one of the mating type–derived motifs are shown in Table 4, along with the remaining genes known to have mating type–specific expression. Genes *MFα1, MFα2, STE3, SAG1, MATα1*, and *MATα2* have mating type α-specific expression. The genes *MFA1, MFA2, STE2, STE6, BAR1, AGA2, MATa1*, and *MATa2* have mating type a-specific expression[31]. The candidate set of coregulated genes corresponding to the P Box-like motif mtα-1 contains 12 genes, including 4 of 14 known mating type–specific genes. Candidate sets derived from those motifs not known to be involved in transcriptional regulation, mta-1 and mta-2, each contained seven ORFs, two of which are mating type–specific. The candidate gene set with matches to either of the Matα2 binding site-like motifs (mta-3 and mtα-mta-1) contains 17 genes—including five of the eight known a-specific genes. That *MATα1* and *MATα2* genes had an upstream match to mtα-mta-1 in their shared intergenic region is consistent with the fact that this locus is repressed by the Matα2p/Mata1p complex in a/α cells[31].

The candidate gene set having upstream sites matching mtα-mta-2—a PRE-like element—contains nine genes, five of which are mating-type regulated. The candidate set for mtα-mta-3 (weakly similar to PRE as noted above) contains 13 genes, two of which are mating-type specific. The gene *STE18* has an upstream match to mtα-mta-3. *STE18* codes for the Gγ subunit of the G protein coupled to both α and a mating factor receptors, and is known to be regulated via the PRE element. Although *STE18* is not thought to have mating type–specific expression, we found the *STE18* transcript to be threefold more abundant in a than α[31]. The candidate gene set corresponding to the Q Box-like motif, mtα-mta-4, contained seven genes, including one a-specific and two α-specific genes. In all, 40 genes were contained in at least one of the candidate sets of coregulated genes described above, and among those 40 were 9 of 14 known mating type–specific genes.

Our experiments showed lower sensitivity to transcripts of low abundance compared with previously published *S. cerevisiae* expression studies[9]. The fraction of all ORFs in the yeast genome that were below detection threshold in both conditions was 70%, 62%, and 68% for galactose vs. glucose, heat shock vs. 30°C, and mating type α vs. a comparisons, respectively. As a result, some transcripts known from previous studies to be differentially expressed between the examined conditions fell below our detection threshold for transcript abundance and thus were not measurably changed in our results. The approach for finding coregulated genes described should become more successful as we optimize experimental methods in whole-genome expression measurement. The methods described here are directly applicable to gene sets clustered by mutant phenotype, by correlated expression over multiple experiments[32], or by other whole-genome methods[45].

Table 4. Change in transcript abundance for mating type α vs. a.

ORF ID	Gene	Change	mtα-1	mta-1	mta-2	mta-3	mtα-mta-1	mtα-mta-2	mtα-mta-3	mtα-mta-4
YPL187W	**MFα1**	**>1.3**	**4**	**-**	**-**	**-**	**1**	**1**	**-**	**1**
YGL089C	**MFα2**	**>1.2**	**1**	**-**	**-**	**-**	**-**	**1**	**-**	**1**
YCL066W	**MATα1**	**>0.7**	**-**	**-**	**-**	**-**	**1**	**-**	**-**	**-**
YCR040W	**MATα1**	**>0.6**	**-**	**-**	**-**	**-**	**1**	**-**	**-**	**-**
YJR004C	**SAG1**	**0.5**	**1**	**-**	**-**	**-**	**-**	**2**	**1**	**-**
YLR040C	**-**	**>0.4**	**-**	**-**	**-**	**-**	**-**	**-**	**1**	**1**
YHR053C	**CUP1**	**0.4**	**1**	**-**	**-**	**-**	**-**	**-**	**-**	**2**
YHR128W	**FUR1**	**0.4**	**-**	**-**	**-**	**-**	**-**	**1**	**-**	**-**
YHR141C	**RPL42B**	**0.4**	**-**	**-**	**-**	**-**	**-**	**-**	**1**	**-**
YGL106W	**MLC1**	**0.4**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**
YGR038W	ORM1	0.3	-	-	-	-	-	1	-	-
YLR355C	ILV5	0.3	-	-	-	-	-	-	-	1
YCR097W	*MATa1*	*ND*	-	-	-	-	-	-	-	-
YCR096C	*MATa2*	*ND*	-	-	-	-	-	-	-	-
YCR039C	*MATα2*	*ND*	-	-	-	-	*1*	-	-	-
YCL067C	*MATα2*	*ND*	-	-	-	-	*1*	-	-	-
YKL178C	*STE3*	*ND*	*1*	-	-	-	-	-	-	-
YKL209C	*STE6*	*ND*	-	-	-	*3*	*3*	-	-	-
YDR301W	YHH1	-0.3	-	-	-	-	-	-	1	-
YLR438W	CAR2	-0.3	1	-	-	-	-	-	-	-
YBR107C	-	-0.3	-	-	-	-	-	-	1	-
YKL080W	VMA5	-0.3	-	-	-	-	-	1	-	-
YML133C	-	-0.3	1	-	-	-	-	-	-	-
YDR342C	HXT7	-0.3	-	-	-	-	-	-	1	-
YLR028C	ADE16	-0.3	1	-	-	-	-	-	-	-
YJR103W	URA8	-0.4	1	-	-	-	-	-	-	-
YML075C	HMG1	-0.4	-	-	-	-	1	-	-	-
YKL128C	PMU1	-0.4	1	-	-	-	-	-	-	-
YCR012W	PGK1	-0.4	1	-	-	-	-	-	-	-
YMR003W	-	-0.4	-	-	1	-	-	-	-	-
YAL061W	-	-0.4	-	-	-	2	2	-	-	-
YPL271W	ATP15	<-0.4	1	-	-	-	-	-	-	-
YBR011C	IPP1	-0.5	-	1	-	-	-	-	-	-
YMR119W	-	-0.5	-	-	-	-	-	-	1	-
YKL214C	-	-0.5	-	-	-	-	-	-	1	-
YKR018C	-	-0.5	-	-	-	1	1	-	-	-
YJR086W	STE18	-0.5	-	-	1	-	-	-	1	-
YBR147W	**-**	**-0.5**	**-**	**2**	**1**	**-**	**-**	**-**	**-**	**-**
YKR071C	**-**	**-0.5**	**-**	**2**	**-**	**1**	**-**	**-**	**2**	**-**
YLL040C	**VPS13**	**-0.6**	**-**	**1**	**2**	**-**	**-**	**-**	**1**	**-**
YJL164C	**SRA3**	**-0.6**	**-**	**1**	**-**	**-**	**-**	**1**	**1**	**1**
YCR024C-A	**PMP1**	**-0.6**	**-**	**-**	**1**	**-**	**-**	**-**	**-**	**-**
YIL015W	**BAR1**	**<-0.7**	**-**	**-**	**2**	**3**	**2**	**-**	**-**	**-**
YFL026W	**STE2**	**<-0.8**	**1**	**-**	**-**	**2**	**2**	**-**	**-**	**1**
YGL032C	**AGA2**	**<-0.9**	**-**	**1**	**-**	**2**	**2**	**2**	**1**	**-**
YNL145W	**MFA2**	**<-1.6**	**-**	**-**	**-**	**3**	**2**	**-**	**-**	**-**
YDR461W	**MFA1**	**-2.0**	**-**	**1**	**1**	**2**	**2**	**1**	**-**	**-**

The top 10 increased or decreased transcripts are in bold type; transcripts not among the top 10 but that had greater than 0.3 absolute change and an upstream match to one of the mating type α and a-related motifs found by AlignACE are in plain text. Transcripts known previously to have mating type–specific expression, but that had an absolute change of less than 0.3, are in italics. Positive change indicates higher transcript abundance in mating type α than a.

Experimental protocol

Strains and growth conditions. *S. cerevisiae* strain FY4 MATa [was used for all growth conditions except as noted. FY4 is a prototroph whose genome is completely sequenced[33,34]. Cultures were grown with aeration in yeast nitrogen base plus ammonium sulfate without amino acids (YNB) at 30°C in a rotary incubator, and supplemented with 2% glucose except as noted. All cultures were harvested in mid–log phase (2–4×10^7 cells/ml) as determined by cell count and dilution plating. Cells were pelleted, quickly washed once with 50 ml dH_2O, frozen in a dry ice–isopropanol bath, and stored at -80°C. For galactose growth, 2% galactose was used in place of glucose. For the heat shocked culture, cells were transferred at mid–log phase to a 39°C shaking water bath. After 13 min the culture had reached 39°C, and the incubation was continued for an additional 20 min. This temperature profile was chosen because heat shock-related transcripts are at maximal levels between 10 and 20 min after reaching the higher temperature[35]. For mating type α cultures, strain FY5—isogenic to FY4 except that it is MATα—was used in place of FY4. All strains were kindly provided by A. Dudley and F. Winston, Harvard Medical School (Boston, MA).

RNA preparation and hybridization. Total cellular RNA was prepared from the frozen cell pellets by hot phenol extraction. Additional phenol and phenol/chloroform/isoamyl alcohol extractions were done before ethanol precipitation. The poly(A) fraction of total cellular RNA was purified by a PolyATract kit (Promega, Madison, WI). The resulting eluate was lyophilized and resuspended in 1–2 μl H_2O. Synthesis, purification, and fragmentation of biotinylated RNA antisense to mRNA transcripts (cRNA) was performed as described[9]. For each culture, expression data were acquired using a set of four chips (yeast antisense A–D; Affymetrix, Santa Clara, CA) designed for expression monitoring of *S. cerevisiae*[9]. Hybridization, washing, and scanning were carried out as described[9] except where noted. Fifteen micrograms of fragmented cRNA in 250 μl buffer were used for all hybridizations, which were carried out either in Affy buffer (Affymetrix) at 40°C or in 6× SSPE-T (0.9 M NaCl, 60 mM NaH_2PO_4, 6 mM EDTA, 0.005% Triton X-100, pH 7.6) at 45°C for 14 to 18 h. For a given chip type, hybridizations were carried out identically for each culture. After hybridization and washing, the chips were stained for 10 min at room temperature with 2 μg/ml streptavidin-phycoerythrin conjugate (SAPE; Molecular Probes, Eugene, OR) in 6× SSPE-T with 1 mg/ml acetylated bovine serum albumin (New England Biolabs, Beverly, MA). Unbound SAPE was removed by rinsing with 6× SSPE-T at 45°C.

Expression data analysis. Perfect match (PM) and single-base mismatch (MM) probe intensities were calculated from raw intensities by the GeneChip software[9] (Affymetrix) For this analysis, genes with MM probes that are perfect matches to a sequence elsewhere within the genome (e.g., *YBL087C* and *YKL006W*) were not considered. Intensities for PM and MM probes were background-subtracted using the average intensity of a set of 36 chip features with consistently low intensity in all of our experiments. For a given chip, PM and MM data were then normalized using the average background-subtracted PM intensity on that chip. This is likely more reliable than normalizing by so-called housekeeping genes (e.g., *ACT1* or *PDA1*) as these can vary between conditions by more than threefold[36]. PM–MM (Δ) is then calculated for each probe pair, and if Δ is less than detection threshold, then Δ is set to this threshold. A detection threshold for PM or MM values on a given chip was chosen to be σ, the standard deviation of background probe intensities on that chip; the threshold for Δ values was then $\sqrt{2}*\sigma$ by propagation of error. The ratio Δ^A/Δ^B of transcript abundance in one condition (A) vs. another (B) was calculated for each corresponding pair of Δ values. Application of the detection threshold prevented unreasonably high (or negative) values for (Δ^A/Δ^B) where a transcript was absent or undetectable in one or both conditions. The change in expression for the transcript corresponding to a given ORF was then calculated as log (median $\{[\Delta^{A1}/\Delta^{B1}], [\Delta^{A2}/\Delta^{B2}],_{,,,} [\Delta^{An}/\Delta^{Bn}]\}$), where n is the number of probe pairs for that ORF. If the median was calculated using a Δ^A/Δ^B value in which the greater of Δ^A and Δ^B was threshold-adjusted, then change in expression for this ORF was not calculated. Otherwise, if the median was calculated using a Δ^A/Δ^B value where either Δ^A or Δ^B were threshold-adjusted, change in expression was stated as being greater than (or less than) the calculated value. The median, as opposed to the mean, was chosen as a measure of central tendency that is robust to outliers. Our rationale for using the median of Δ ratios rather than the ratio of Δ medians was that, while the magnitude of Δ for features of different sequence can vary considerably, Δ^A/Δ^B should be less variant. An alternative measure of expression change (not used here) is fractional change: median $([\Delta^{A1}-\Delta^{B1}]/[\Delta^{A1}+\Delta^{B1}], [\Delta^{A2}-\Delta^{B2}]/[\Delta^{A2}+\Delta^{B2}],_{,,,} [\Delta^{An}-\Delta^{Bn}]/[\Delta^{An}+\Delta^{Bn}])$. This measure should be robust to cases where a transcript is below detection threshold in one condition but not the other. Optimality of these or other methods[9] for intensity data analysis have yet to be demonstrated experimentally.

Finding sites that match motifs and calculating motif occurrence score. Sites were scored against motifs using the method of Berg and von Hippel[37]. To define a threshold score for a good matching site, we first calculated the mean score μ and standard deviation σ for the set of sites that were aligned by AlignACE. We determined the occurrence score—the estimated fraction of genes with upstream regions bound by the (putative) protein corresponding to a given motif—in the following way: We found the fraction of genes with upstream matches, defining a matching site as one with a Berg-von Hippel score greater than or equal to μ. The occurrence score was then calculated as twice this fraction, as half of "real" sites will have scores better than μ if "real" sites have a symmetric score distribution. For discovery of specific upstream sites, the threshold score for a "good" match was lowered to μ - 3σ to avoid missing potential binding sites.

Modifications to GMS. A number of alterations were made to the GMS algorithm[15]. (1) Simultaneous multiple motif searching was replaced with an iterative masking approach, allowing a more efficient search for subtle motifs. (2) Both strands of DNA were considered, so that when a potential site was examined, either the site or its reverse complement—but not both—were

added to the alignment. (3) Scoring of alignments was by frequency of occurrence in the intergenic DNA sequence of *S. cerevisiae* using the method described above. (4) Near-optimum sampling method was improved so that it tended to result in higher scoring alignments and so that all columns spanned by the initially chosen columns were considered. (5) The model for base frequencies of nonsite sequence was fixed using the background nucleotide frequencies of *S. cerevisiae*. (6) The code was rendered portable to DEC Unix and Windows platforms, in addition to Silicon Graphics and Sun Unix systems.

AlignACE settings. AlignACE was used with the following settings: initial alignment used a column-sampling approach with 10 columns; the expected number of sites was 10; maximum number of initial sampling runs was 500; iterative masking to find multiple motifs was performed a maximum of 100 times; near-optimum sampling commenced after 50 consecutive sampling runs without an increase in alignment score; iterative masking was terminated after three consecutive cases of nonconvergence or nonpositive alignment score. AlignACE implements a scoring method for "goodness" of alignment that has previously been described by equation 10 of ref. 38. AlignACE and all other software written for this work is available as well as the mRNA quantitation data used in this study".

Measuring similarity between DNA motifs. To identify described motifs that might be similar to the newly identified motif, the relevant literature was searched. Additionally, the TRANSFAC Release 3.2 was searched[40] using PatternSearch or MatInspector 2.1 (ref. 41) with a threshold of 85% identity for "core" nucleotides and 60% for overall identity. To assess similarity more quantitatively once a putatively similar motif was identified, DNA site weight matrices[37] were examined pairwise in all possible alignments. The alignment that minimizes the sum of squared differences between matrix elements was chosen. When necessary, as in the case of imperfectly overlapping matrices, matrix elements were taken as their expectation value, based on the base composition of *S. cerevisiae*. If aligned matrices did not overlap by more than 5 bases, they were considered not similar. To further compare a given matrix A with a matrix B, a submatrix of A defined by the region of overlap between aligned A and B matrices was constructed. Each site used in generating the A and B matrices was scored against submatrix A. Student's *t*-test score was calculated from the set of A site scores and the set of B site scores. Matrix A was said to detect matrix B if *t* was above a threshold (described below). The procedure was repeated using a submatrix of B, and matrix A was then said to be similar to matrix B if either matrix detected the other. The threshold for *t* was obtained using a negative control set of 14 literature-derived matrices that each bind different proteins. The false-positive rate for matrix similarity was calculated to be the number of similar pairs in the negative control set divided by 91 (the number of pairwise comparisons). The value of *t* corresponding to a 5% false-positive rate was found to be 3.26 by linear interpolation.

Acknowledgments

We thank A. Dudley, F. Winston, K. Robison, A. Neuwald, M. Temple, T. Smith, T. Schneider, P. Johnson, T. Blackwell, M. Johnston, and K. Struhl for their advice and comments. We thank our collaborators at Affymetrix—particularly D. Lockhart, J. Warrington, L. Wodicka, and R. Blalock—for their support and enthusiasm. Members of the Church lab contributed invaluable discussions and critical review. F.R. and J.H. were each supported by an NSF Graduate Fellowship. G.M.C. was an investigator with the Howard Hughes Medical Institute. This work was generously supported by the US Department of Energy (grant no. DE-FG02-87-ER60565), the Office of Naval Research (grant no. N00014-97-1-0865), and the Lipper Foundation. Use of an SGI Power ChallengeArray was provided by the NCSA at the University of Illinois at Urbana-Champaign.

1. Pennisi, E. 1997. Laboratory workhorse decoded. *Science* **277:**1432–1434.
2. Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T,. Burland, V., Riley, M. et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:**1453–1474.
3. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldman, H. et al. 1996. Life with 6000 genes. *Science* **274:**563–567.
4. Chen, P., Ailion, M., Bobik, T., Stormo, G., and Roth, J. 1995. Five promoters integrate control of the cob/pdu regulon in *Salmonella typhimurium*. *J. Bacteriol.* **177:**5401–5410.
5. Chuang, S.E., Daniels, D.L., and Blattner, F.R. 1993. Global regulation of gene expression in *Escherichia coli*. *J. Bacteriol.* **175:**2026–2036.
6. Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270:**467–470.
7. Lockhart, D.J., Dong, H.L., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S. et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14:**1675–1680.
8. DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M. et al. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14:**457–460.
9. Wodicka, L., Dong, H., Mittmann, M., Ho, M.-H., and Lockhart, D.J. 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **15:**1359–1366.
10. Muhlrad, D., Decker, C.J., and Parker, R. 1995. Turnover mechanisms of the stable yeast PGK1 mRNA. *Mol. Cell. Biol.* **15:**2145–2156.
11. Jacobson, A. and Peltz, S.W. 1996. Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. *Annu. Rev. Biochem.* **65:**693–739.
12. Lipman, D.J. 1997. Making (anti)sense of non-coding sequence conservation. *Nucleic Acids. Res.* **25:**3580–3583.
13. Frech, K., Quandt, K., and Werner, T. 1997. Software for the analysis of DNA sequence elements of transcription. *CABIOS* **13:**89–97.
14. Bailey, T.L. and Elkan, C. 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning Journal* **21:**51–83.
15. Neuwald, A.F., Liu, J.S., and Lawrence, C.E. 1995. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.* **4:**1618–1632.
16. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262:**208–214.
17. Lohr, D., Venkov, P., and Zlatanova, J. 1995. Transcriptional regulation in the yeast GAL gene family: a complex genetic network. *FASEB J.* **9:**777–787.
18. Schneider, T.D. and Stephens, R.M. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18:**6097–6100.
19. Freeman, K.B., Karns, L.R., Lutz, K.A., and Smith, M.M. 1992. Histone H3 transcription in *Saccharomyces cerevisiae* is controlled by multiple cell cycle activation sites and a constitutive negative regulatory element. *Mol. Cell. Biol.* **12:**5455–5463.
20. Osley, M.A. 1991. The regulation of histone synthesis in the cell cycle. *Annu. Rev. Biochem.* **60:**827–861.
21. Breeden, L. 1996. Start-specific transcription in yeast. *Curr. Top. Microbiol. Immunol.* **208:**95–127.
22. McInerny, C.J., Partridge, J.F., Mikesell, G.E., Creemer, D.P., and Breeden, L.L. 1997. A novel Mcm1-dependent element in the SWI4, CLN3, CDC6, and CDC47 promoters activates M/G1-specific transcription. *Genes Dev.* **11:**1277–1288.
23. Herskowitz, I., Rine, J., and Strathern, J. 1992. Mating-type determination and mating-type interconversion in *Saccharomyces cerevisiae*, pp. 583–656, in *Gene expression*, Vol. 2. Jones, E.W., Pringle, J.R., and Broach, J.R. (eds.). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
24. Johnston, M. and Carlson, M. 1992. Regulation of carbon and phosphate utilization, pp. 193–281 in *Gene expression*, Vol. 2. Jones, E.W., Pringle, J.R., and Broach, J.R. (eds.). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
25. Craig, E.A. 1992. The heat-shock response of *Saccharomyces cerevisiae*, pp. 501–537, in *Gene expression*, Vol. 2. Jones, E.W., Pringle, J.R., and Broach, J.R. (eds.). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
26. Schmitt, A.P. and McEntee, K. 1996. Msn2p, a zinc finger DNA-binding protein, is the transcriptional activator of the multistress response in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* **93:**5777–5782.
27. Rowley, A., Johnston, G.C., Butler, B., Werner-Washburne, M., and Singer, R.A. 1993. Heat shock-mediated cell cycle blockage and G1 cyclin expression in the yeast *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **13:**1034–1041.
28. Vashee, S., Xu, H., Johnston, S.A., and Kodadek, T. 1993. How do "Zn2 cys6" proteins distinguish between similar upstream activation sites? Comparison of the DNA-binding specificity of the GAL4 protein in vitro and in vivo. *J. Biol. Chem.* **268:**24699–24706.
29. Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T. et al. 1998. SGD: saccharomyces genome database. *Nucleic Acids Res.* **26:**73–79.
30. Ni, H.T. and LaPorte, D.C. 1995. Response of a yeast glycogen synthase gene to stress. *Mol. Microbiol.* **16:**1197–1205.
31. Sprague, G.F. and Thorner, J.W. 1992. Pheromone response and signal transduction during the mating process of *Saccharomyces cerevisiae*, pp. 657–744, in *Gene expression*, Vol. 2. Jones, E.W., Pringle, J.R., and Broach, J.R. (eds.). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
32. Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L. et al. 1998. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci USA* **95:**334–339.
33. Winston, F., Dollard, C., and Ricupero-Hovasse, S.L. 1995. Construction of a set of convenient *Saccharomyces cerevisiae* strains that are isogenic to S288C. *Yeast* **11:**53–55.
34. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B. et al. 1996. Life with 6000 genes. *Science* **274:**563–567.
35. Miller, M.J., Xuong, N.H., and Geiduschek, E.P. 1982. Quantitative analysis of the heat shock response of *Saccharomyces cerevisiae*. *J. Bacteriol.* **151:**311–327.
36. Wenzel, T.J., Teunissen, A.W., and de Steensma, H.Y. 1995. PDA1 mRNA: a standard for quantitation of mRNA in *Saccharomyces cerevisiae* superior to ACT1 mRNA. *Nucleic Acids Res.* **23:**883–884.
37. Berg, O.G. and von Hippel, P.H. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193:**723–750.
38. Liu, J.S., Neuwald, A.F., and Lawrence, C.E. 1995. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Stat. Assoc.* **90:**1156–1170.
39. http://arep.med.harvard.edu/mrnadata
40. Wingender, E., Kel, A.E., Kel, O.V., Karas, H., Heinemeyer, T. et al. 1997. TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation. *Nucleic Acids Res.* **25:**265–268.
41. Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. 1995. MatInd and matInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* **23:**4878–4884.
42. Simon, J.A. and Lis, J.T. 1987. A germline transformation analysis reveals flexibility in the organization of heat shock consensus elements. *Nucleic Acids Res.* **15:**2971–2988.
43. Schuller, C., Brewster, J.L., Alexander, M.R., Gustin, M.C., and Ruis, H. 1994. The HOG pathway controls osmotic regulation of transcription via the stress response element (STRE) of the Saccharomyces cerevisiae CTT1 gene. *EMBO J.* **13:**4382–4389.
44. Martinez-Pastor, M.T., Marchler, G., Schuller, C., Marchler-Bauer, A., Ruis, H. et al. 1996. The *Saccharomyces cerevisiae* zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). *EMBO J.* **15:**2227–2235.
45. Tavazoie, S. and Church, G.M. 1998. Quantitative whole-genome analysis of DNA-protein interactions by in vivo methylase protection in *E. coli*. *Nat. Biotechnol.* **16:**566–571.