

Gene Expression From Random Libraries of Yeast Promoters

Martin Ligr,* Rahul Siddharthan,[†] Fredrick R. Cross* and Eric D. Siggia*¹

*The Rockefeller University, New York, New York 10021 and [†]Institute of Mathematical Sciences, Taramani, Chennai 600113, India

Manuscript received October 20, 2005
Accepted for publication January 3, 2006

ABSTRACT

Genomewide techniques to assay gene expression and transcription factor binding are in widespread use, but are far from providing predictive rules for the function of regulatory DNA. To investigate more intensively the grammar rules for active regulatory sequence, we made libraries from random ligations of a very restricted set of sequences. Working with the yeast *Saccharomyces cerevisiae*, we developed a novel screen based on the sensitivity of ascospores lacking dihydroxyacetone to treatment with lytic enzymes. We tested two separate libraries built by random ligation of a single type of activator site either for a well-characterized sporulation factor, Ndt80, or for a new sporulation-specific regulatory site that we identified and several neutral spacer elements. This selective system achieved up to 1:10⁴ enrichment of the artificial sequences that were active during sporulation, allowing a high-throughput analysis of large libraries of synthetic promoters. This is not practical with methods involving direct screening for expression, such as those based on fluorescent reporters. There were very few false positives, since active promoters always passed the screen when retested. The survival rate of our libraries containing roughly equal numbers of spacers and activators was a few percent that of libraries made from activators alone. The sequences of ~100 examples of active and inactive promoters could not be distinguished by simple binary rules; instead, the best model for the data was a linear regression fit of a quantitative measure of gene activity to multiple features of the regulatory sequence.

IN spite of the impressive technologies available for assaying gene expression and protein localization, and the availability of related genomes, the prediction of expression from sequence is still very imprecise (SIGGIA 2005). Even in the favorable case of Gal4-regulated genes where the protein localization and genetic data agree quite well, there remain many non-functional sites in the genome (REN *et al.* 2000). There is also considerable variation between labs (IYER *et al.* 2001; SIMON *et al.* 2001), and the geneticist's assay for function, change in response to gene deletion, may not agree with the protein localization data (BEAN *et al.* 2005). So one may ask, (1) Does a protein bound to DNA confer regulation?, (2) How important is the actual core promoter for the interpretation/integration of the signal from bound transcription factors?, and (3) How important is the sequence environment to the activity of specific binding sites (SEKINGER *et al.* 2005)?

One measure of the complexity of a process, *e.g.*, the mapping from sequence to expression, is the length in bits of the most compact rule required to define it. An upper bound is simply a spreadsheet of expression data under all conditions. The number of ways of arranging the 600 bases in the typical yeast promoter is unimaginably large and we presume that it is only the binding of

a limited number of factors that matter. Even so, the number of possible combinations again far exceeds the number of species one could hope to sequence for comparative genomics.

An alternative strategy for exploring the richness of the mapping from sequence to expression is to assay libraries of synthetic promoters. An important branch of computer science deals with algorithms that learn from examples and incidentally has to control for the tendency to construct overly specific rules, given the generally limited number of examples in the training set. The simplest types of learning environments provide the algorithm with a random set of positive and negative examples of the unknown rule and do not allow the "learner" to query the "oracle" about an example of its choosing. This is precisely the situation realized by screening a random library for function. Specifically, we chose a limited number of binding sites ("words") and asked whether random combinations lead to expression (yield a meaningful "sentence").

We have built our assay around sporulation in yeast. Sporulation is a cellular differentiation process that is triggered when diploid (*MATa/MATα*) budding yeast are subjected to nitrogen starvation in the absence of fermentable carbon sources. The cell exits the cell cycle and completes a single round of DNA replication. Homologous chromosomes pair, recombine, and undergo two meiotic divisions in the nucleus. At the spindle-pole body, a formation of prospore wall is

¹Corresponding author: Center for Studies in Physics and Biology, Rockefeller University, Box 25, 1230 York Ave., New York, NY 10021.
E-mail: siggiae@mail.rockefeller.edu

initiated, leading to engulfment of haploid meiotic products and cytoplasmic material in four prospores. Finally, several layers of spore-wall material are deposited on the surface of prospores, giving rise to four ascospores, enclosed in an ascus formed by the cell wall of the vegetative cell (ESPOSITO and KLAPHOLZ 1981). The sporulation-specific genes are expressed in ordered sequence and on the basis of the timing can be divided into four classes: early, middle, middle-late, and late sporulation genes (MITCHELL 1994). The onset of the middle phase of sporulation depends on successful completion of recombination and segregation of homologous chromosomes at the end of prophase I. An important regulator of the middle phase is Ndt80, which is a target of the pachytene checkpoint (HEPWORTH *et al.* 1998). Over 70% of genes induced in this stage contain in their upstream regulatory sequence (URS) a middle sporulation element (MSE) (CHU *et al.* 1998), which is a binding site for Ndt80. Many of the genes expressed in the middle-late phase are involved in spore-wall assembly, and about half of them contain MSE (CHU *et al.* 1998).

Sporulation is a good environment for our assay since it has been well studied genetically, and it has been subject of two microarray experiments, although with considerable variation between them (CHU *et al.* 1998; PRIMIG *et al.* 2000). There are two primary DNA-binding regulatory factors, Ume6/Ime1 and Ndt80 (CHU *et al.* 1998; PRIMIG *et al.* 2000), so the “vocabulary” available from which to construct promoters is small. [In addition, the Sum1 repressor competes with Ndt80 for binding to an extended site (PIERCE *et al.* 2003; JOLLY *et al.* 2005) and the general factor Abf1 is required for some meiotic genes (KASSIR *et al.* 2003)]. Sporulation is a terminal process, so while there is some cell-to-cell or strain-to-strain variation, the end point is the same, making timing not so crucial. We screened by assaying for a tough spore wall in a strain deleted for an enzyme essential for wall formation, *DIT1*, which was placed under the control of our promoter library. A gene coding for a sporulation-specific enzyme should have simple, precise regulation. There is a plausible link between fitness and the phenotype that we assay for; therefore variation in expression should matter. The combination of sporulation, destruction of unprotected spores, and germination is also very sensitive. It allows for the detection of rare functional strains with few false positives; *i.e.*, upon retesting they “breed true.” This would not be possible with a fluorescent marker because of background and low throughput (MILLER and WIDOM 2003).

Regulatory sites are commonly formed into tandem arrays and then assayed in a reporter construct. Natural promoters are almost never direct repeats and we expected that synthetic promoters built with spacers could be modeled by simply counting activators. Instead, we found that the cell generates very different outputs from very similar promoters. The data supporting this assertion are the subject of this article.

MATERIALS AND METHODS

Strains and media: Strains YL332 (*MATa ura3 dit1::kanMX*), YL334 [*MAT α ura3 dit1::kanMX*], and YL344 [*MATa/MAT α ura3/ura3 dit1::loxP/dit1::loxP*] were used in our experiments and cultured in standard media (SHERMAN 1991).

Vectors: The reporter vector was based on pRS416 [*URA3/CEN6*] and contained a nonfunctional *DIT2* fragment, *Bam*HI site, *MEL1* TATA region, GFP-*DIT1* fusion, and *DIT1* terminator (Figure 1).

Promoter library construction and screening: An activator sequence, four species of inert random sequences, and two terminal adaptors (supplemental Table S1 at <http://www.genetics.org/supplemental/>) were ligated and cotransformed into YL344 together with a linearized reporter vector. The transformants were collected and induced to sporulate in liquid medium. The samples of sporulated cultures were treated with glusulase (Perkin-Elmer Life Sciences, Boston) and plated onto YEPD plates to isolate surviving spores. Survivors were mated with an isogenic strain of complementary mating type and the diploids were collected. Total genomic DNA was isolated from these clones, amplified, and sequenced.

Dityrosine test: The test for presence of dityrosine in spore walls was performed as described by BRIZA *et al.* (1990).

Diethylether survival: The strains were patched and sporulated as for a dityrosine test, but each set of patches was sporulated in triplicate. Membranes carrying sporulated cells were floated on diethylether for 5, 10, and 20 min. After brief drying, the membranes were replica plated onto YPD plates and regrowth of patches was scored after 2 days. For further details, see supplementary materials at <http://www.genetics.org/supplemental/>.

RESULTS

Selection of active elements: *DIT1* and *DIT2* are expressed during the mid-late period of sporulation, and in their shared upstream region (Figure 1) a common variant, GTCGCAAAA, of the MSE, GNCA CAAAA, binds the sporulation-specific transcription factor Ndt80 (FRIESEN *et al.* 1997). Although this element is functional *in vitro* (LAMOUREUX *et al.* 2002), the regulation of *DIT1* is not fully understood and is a good illustration of how difficult it is to localize regulation to specific binding sites (see supplemental methods at <http://www.genetics.org/supplemental/>). For our library, we took 27 bp surrounding this MSE site and called it MSE-DIT.

We also computationally searched the common *DIT1*,2 regulatory region in *Saccharomyces cerevisiae* and related species (see supplemental methods at <http://www.genetics.org/supplemental/>) for other overrepresented motifs. We found a motif with consensus TRAGGGY, which in addition was overrepresented in the mid-late and late genes common to the two expression array experiments (CHU *et al.* 1998; PRIMIG *et al.* 2000). We chose our second activator element to be a 27-bp sequence bracketing one of these sites and termed it SSE-DIT sporulation stress-like element because of its similarity to the stress element AGGGG (MARTINEZ-PASTOR *et al.* 1996).

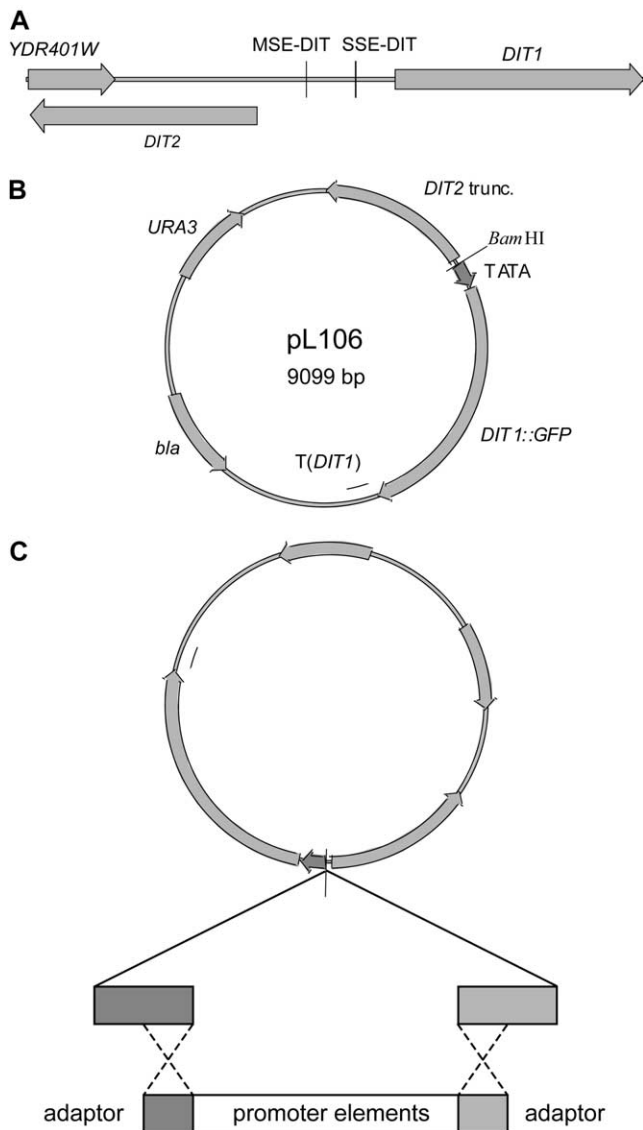


FIGURE 1.—(A) Map of *DIT1/DIT2* genomic region. (B) Map of reporter vector. (C) Scheme of homologous recombination reaction used to insert artificial promoter sequences into the reporter vector. Dashed lines indicate homologous recombination between *in vitro*-generated artificial promoter sequence and the corresponding regions in the cut vector. T(*DIT1*), terminator region of *DIT1* gene; *DIT2* trunc, truncated, nonfunctional fragment of *DIT2* gene; *bla*, β -lactamase.

Screen design and validation: To study constraints placed on architecture of regulatory DNA sequences, we devised a screen to select for artificial sequences that have the capacity to drive expression of late sporulation genes in *S. cerevisiae*. We took advantage of the role of the gene *DIT1* in maturation of ascospores: Dit1 catalyzes the reaction leading from L-tyrosine to the tyrosine-containing intermediate product, while Dit2, transcribed divergently from *DIT1*, is responsible for the following dimerization reaction leading to the dityrosine-containing precursors (BRIZA *et al.* 1994). These are incorporated into the top layer of ascospore walls

(BRIZA *et al.* 1986). Dityrosine renders the spores resistant to lytic enzymes, high temperature, and diethylether (BRIZA *et al.* 1990). Using *DIT1* as a reporter gene, we could exploit the sensitivity of spores lacking dityrosine to these treatments. We constructed a reporter vector (Figure 1) that contained *DIT1* tagged with GFP at the N terminus; a minimal promoter unable to drive expression without URS (MELCHER *et al.* 2000); and an inactive fragment of *DIT2* to insulate the minimal promoter from spurious transcription factor binding sites elsewhere on the vector. To test the functionality of the *DIT1::GFP* reporter, we inserted the *DIT1/DIT2* intergenic region into the *Bam*HI site of the vector. When we compared this construct to a similar plasmid carrying the wild-type *DIT1/DIT2* locus, we did not detect any difference in glusulase sensitivity or dityrosine fluorescence levels, indicating that our reporter construct is fully functional.

In vitro-generated promoter sequences (constructed as described in MATERIALS AND METHODS and below) were cotransformed into *dit1/dit1* cells together with linearized reporter vector (Figure 1C), and cells containing repaired vector were selected for uracil prototrophy. To check the extent of gap repair of the cut plasmid, we compared the colony counts after transforming the cells with cut plasmid only, and cut plasmid plus library fragments, and observed at least a 25-fold increase in colony number from inclusion of the library fragments. The diploids were induced to sporulate and then treated with glusulase, a mix of lytic enzymes targeting inner layers of ascospores. Surviving spores, which presumably contained a synthetic URS capable of driving expression of a functional level of *DIT1*, were allowed to germinate on complete rich medium. The haploid colonies were mated with cells of complementary mating type and the diploids were collected, induced to sporulate, and tested for dityrosine fluorescence and diethylether resistance.

After treatment with glusulase (Figure 2, A and B), 1 in 10^4 cells carrying empty vector survived, relative to survival of the cells carrying a positive control plasmid with the native *DIT1/DIT2* regulatory region. To test reproducibility of selection on the basis of dityrosine fluorescence of ascospores, we isolated library plasmids showing varying levels of fluorescence, transformed them into the original plasmid-free strain, and compared the levels of dityrosine fluorescence after sporulation (Figure 2C). The differences between the original transformants and the retransformants were $<5\%$. We also performed a double-blind test on complete libraries described below to assess reproducibility of dityrosine fluorescence levels when assayed in large-scale, temporally distant experiments. Again, the fluorescence levels obtained agreed within 5%.

When we subjected the transformants obtained by glusulase selection to a second glusulase treatment, they exhibited a level of resistance equivalent to the positive

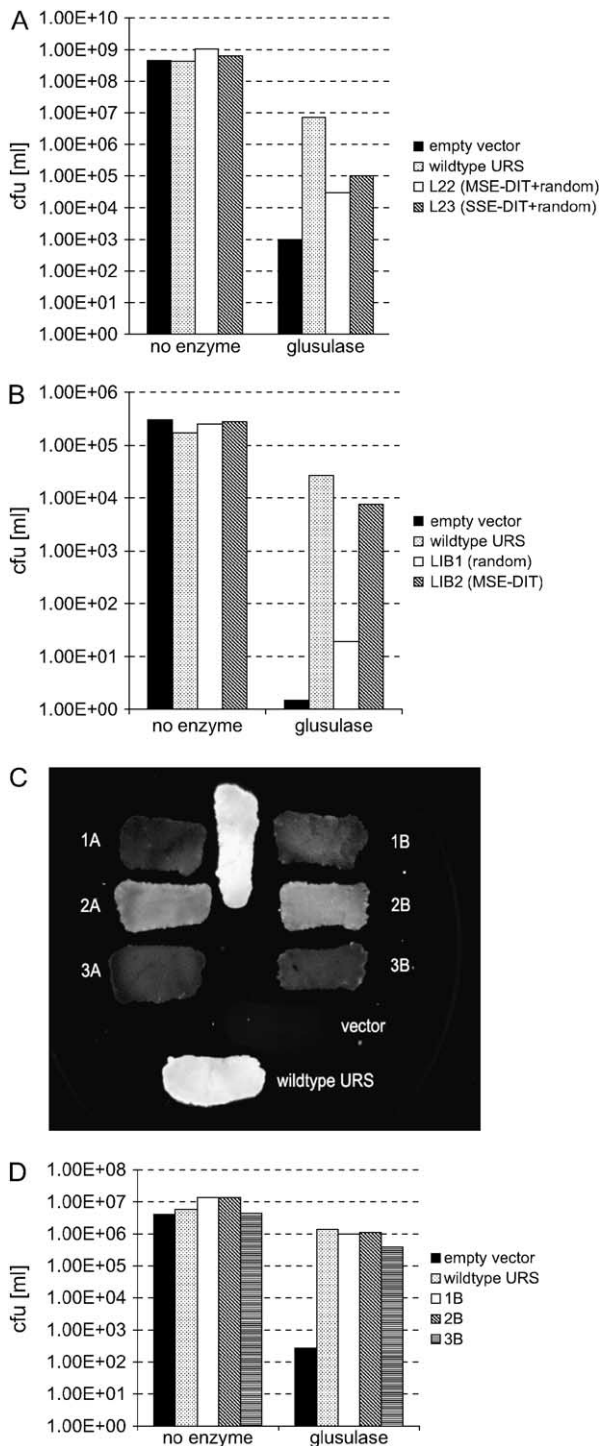


FIGURE 2.—(A and B) Survival rates of libraries after screening. (C) Reproducibility of dihydroxyacetone phosphate fluorescence. Library plasmids conferring glusulase resistance to three different clones (1A, 2A, 3A) were recovered and transformed into the original plasmid-free tester strain (1B, 2B, 3B), and compared to strains carrying empty vector and the positive control with wild-type URS for dihydroxyacetone phosphate fluorescence after sporulation. (D) Resistance of strains 1B, 2B, and 3B to glusulase treatment.

control, irrespective of their different fluorescence levels. Thus the glusulase treatment appeared to be essentially a threshold selection for a minimal level of *DIT1* expression, without any differential survival dependent on expression above the threshold.

Library building and selection: We first constructed libraries consisting only of activators (length 31 bp, including 4-bp sticky ends) (supplemental Table S1 at <http://www.genetics.org/supplemental/>), ligated with random orientations. For library L2, constructed from the MSE-DIT site, approximately one in three of the clones survived the treatment with glusulase, relative to the strain with wild-type URS (Figure 2B). Forty-one percent of L2 clones before selection had dihydroxyacetone phosphate fluorescence levels >10% of the positive control containing the wild-type URS. Library L15 containing the SSE-DIT site showed even better survival rates (76% of the positive control), and 85% of clones before selection showed fluorescence levels >10% of the positive control.

Libraries L22 (MSE-DIT + random) and L23 (SSE-DIT + random) were created from the same active elements as above, supplemented with a fourfold excess of random spacers. There were four spacer elements (each composing 20% of the ligation mix), all 24 bp long (including the same 4-bp sticky ends) (supplemental Table S1 at <http://www.genetics.org/supplemental/>). These elements were designed by selecting random sequences constrained by the AT/GC ratio of 0.6 typical for yeast noncoding regions. The sequences were also screened for known and predicted transcription factor binding sites. The synthetic promoter elements consisting solely of the random elements were essentially transcriptionally inactive (Figure 2B, library L1).

After the screening with glusulase, ~1 in 100 clones carrying the library sequences L22 and L23 survived relative to the positive control (Figure 2A), *i.e.*, a factor of 30–100 lower than the proportion of survivors from libraries without spacers. We then sequenced ~100 clones from the unselected and selected pools for both spacer-containing libraries and subjected them to further tests.

Before selection, the L22 pool contained 2% of clones that after sporulation showed dihydroxyacetone phosphate fluorescence >10% of wild-type value; after selection by glusulase treatment, 94% of the clones were fluorescent after sporulation. In library L23, 8% of clones produced spores with dihydroxyacetone phosphate fluorescence >10% before selection. These clones were all resistant to diethylene glycol and survived glusulase treatment as wild type when tested individually. (Thus 10% of wild-type fluorescence apparently represents a level of *DIT1* expression sufficient for survival.) After selection 98% of the L23 clones had >10% fluorescence. The maximum fluorescence in L23 was 88% of wild type *vs.* only 35% of wild type for L22. It therefore appears that L23 contained a higher proportion of functional clones, consistent with the survival data in Figure 2A. The fluorescence test was done a second time

TABLE 1

Elimination of clones during individual selection steps

Preselection	Initial count	10% fluor 1	10% fluor 2	Ether
Library L22	97	0	2	2
Library L23	87	6	7	7
Postselection	Screen	10% fluor 1	10% fluor 2	Ether
Library L22	99	93	90	90
Library L23	112	110	110	110

“Initial count” is the number of clones randomly selected from the library before selection; “screen” is the number of clones selected after the glusulase screen; “10% fluor 1” is the number of clones left after applying a 10% fluorescence threshold of dityrosine fluorescence; “10% fluor 2” is the number of clones left after the second independent round of measurements of dityrosine fluorescence and applying 10% fluorescence threshold, followed by a diethylether screen (“ether”) on the survivors.

for all four sequenced pools, and all positives were confirmed again by testing with ether (Table 1).

SSE-DIT drives sporulation-specific expression:

When active concatenates of the SSE-DIT site from L23 were placed upstream of the *DIT1::GFP* reporter, green fluorescence was observed in ascii roughly at the time of appearance of the outline of individual spores (Figures 3 and 4). On synthetic medium with glucose-

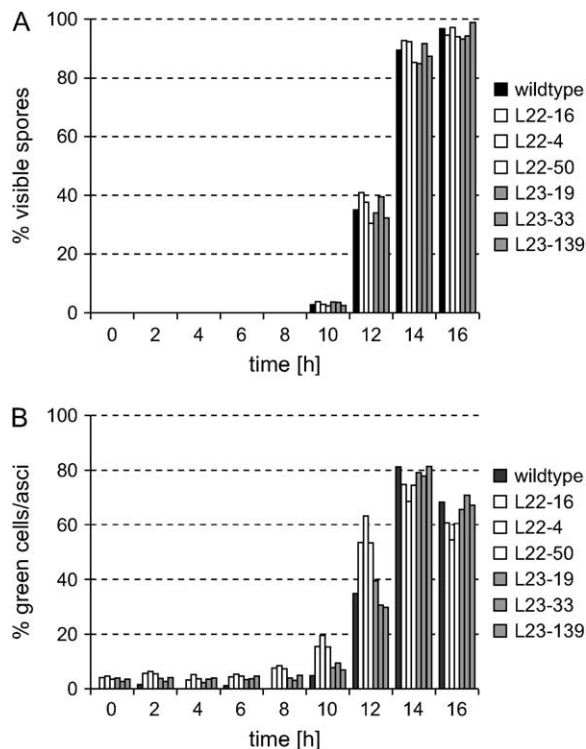


FIGURE 3.—(A) Percentage of ascii-containing spores in the course of sporulation. (B) Percentage of cells/ascii showing green fluorescence. The cells were grown in SC-ura medium, washed, and incubated in 2% potassium acetate. No green fluorescence was seen in control cells with an empty vector.

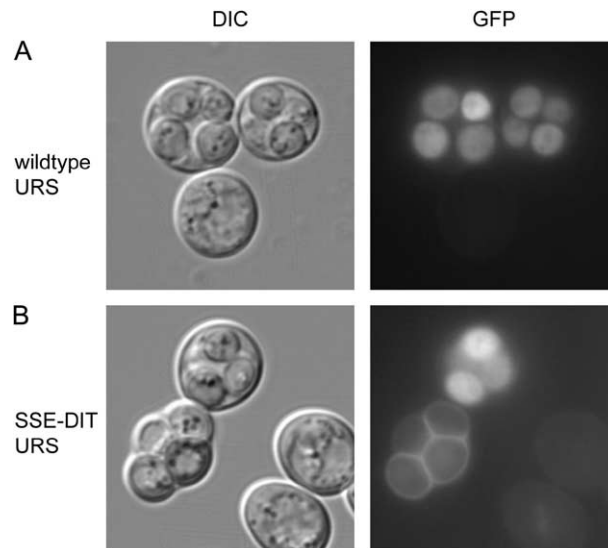


FIGURE 4.—*DIT1::GFP* under control of an active SSE-DIT promoter is visible only in the late phase of sporulation. (A) Cells harboring GFP reporter under the control of native promoter (wild-type URS). (B) Cells harboring GFP reporter under the control of SSE-DIT promoter (clone L23-19). The samples were taken 12 hr after the induction of sporulation. Field of view was selected to contain cells both before and after the appearance of spore outlines.

supporting vegetative growth, the green signal was visible in $3.4\% \pm 0.6\%$ of diploid cells. The same applies to the MSE-DIT element containing the predicted Ndt80-binding site ($4.0\% \pm 0.6\%$). Thus expression driven by either SSE-DIT or MSE-DIT was sporulation specific.

Stress conditions (heat shock at 42° and osmotic shock with 1 M NaCl) applied to the same strains as Figure 3 did not increase the GFP signal above the level observed in vegetatively growing cells (data not shown). During growth on complete medium containing acetate, $10.0\% \pm 2.1\%$ cells containing active SSE-DIT showed GFP signal ($7.1\% \pm 4.1\%$ cells with MSE-DIT). Thus the SSE-DIT element is only highly active during late sporulation and not in stressed or unstressed vegetative cells. It exhibits sporulation specificity comparable to the known sporulation-specific MSE element.

Analysis of sequences: Prior to examining the sequences we excluded a few members of the unselected pools with fluorescence $>10\%$ of wild type and a few selected clones with fluorescence $<10\%$ (Table 1). (This was to ensure that fortuitous positive clones picked from the unselected library, or poor expressors that leaked through the selection system near the threshold for *DIT1* expression, did not confound the final results.) The remaining sets of sequences (low expressors preselection and high expressors postselection) constitute our high confidence set (supplemental Tables S2 and S3 at <http://www.genetics.org/supplemental/>).

We first examined the two sets visually to see if any simple rules were discernible, and if there were similar clones in the active and inactive libraries that would

TABLE 2

Contrasted samples of sequenced clones in the inactive and active libraries with generally similar structures, and rationalization of differences in activity

Clone	Element arrangement	Fluorescence	
	A = MSE-DIT		
L22K-9	A><A	0.06	
L22K-53	A><A	0.08	
L22-105	A><A A><A	0.24	Copy number matters
L22K-102	A> A><A	0.08	
L22-47	A><A <A	0.22	Reverse complement matters
	A = SSE-DIT		
L23K-78	<A A>	0.08	
L23-9	1> 3><A A>	0.42	Add random elements
L23-137	3><1 <A A><2	0.18	
L23K-73	<A <A	0.03	
L23-52	<A <A <2 3>	0.14	Add downstream elements
L23K-55	A> A><A	0.09	
L23-49	4> A> A><A	0.26	Add upstream element
L23K-70	A><2 A>	-0.03	
L23-29	A><4 A>	0.22	Which element matters

See supplemental Tables S2 and S3 for complete data for all sequenced clones. “Fluorescence” refers to di-tyrosine fluorescence of spores, relative to wild-type. All “LXXX” elements are inactive preselection clones.

either point to errors or suggest subtleties in the rules that could distinguish one set from the other. We first consider the data as binary since under the screen the survival of positive clones is very high when retested and the survival of the negative ones correspondingly low. (The more complex alternative is to correlate the di-tyrosine fluorescence with the sequence, treating it as a real number.) The sequences are characterized by pairs of symbols denoting the elements (A for activator, 1–4 for the random element) and their orientation (>, <), with a vertical bar for delineation. The active *DITI* gene always lies to the right (downstream) of the promoter. Table 2 shows a few cases that exhibit interesting patterns, but the following discussion includes all available data (*e.g.*, supplemental Tables S2 and S3 at <http://www.genetics.org/supplemental/>) and thus is nonanecdotal within the limits of our data.

There is an excess of shorter elements in the inactive MSE-DIT pool (L22K) (7 *vs.* 1 length 2, and 13 *vs.* 7 length 3). The only length 2 functional element <A|4> (L22-61) had a single activator and random element (whereas several elements with two activators were found in the nonfunctional set). Random element 4 never occurred among the length 2 clones in L22K, but this is not significant because of our sample size. Among longer inactive clones, there are eight other occurrences of A followed by 4, but always in the >> or << orientation. However, the active combination <A|4> does occur in the six-element clone L22K-36 in combination with

other activator sites and is inactive. So we might conclude that there is something about the A-4 boundary and <> orientation that favors activity. But within L22, the A-4 combination does not display an orientation bias.

If we instead search for occurrences of the inactive length 2 clones in the active library, we see that although L22K-9,53, |A><A|, is inactive, adding an additional activator in either orientation to the downstream end restores activity (L22-7,47). Furthermore, a dimer of L22K-53, L22-18 is active. There is no puzzle here if we focus on the fluorescence, since the active clones are only two to three times brighter than the inactive ones and the selection is rather sharp around a fluorescence of ~0.1. Among length 4 clones consisting solely of activators, there are 9 inactive and only 1 active. The active clone has the orientation ><><, as already noted, but it is not surprising that with 16 possible orientations of four activators that a particular one is not found in a sample of nine ($P \sim 0.5$). The ><>< combination of activators was found three times (with a downstream |A> added) among the length 5 active clones. In general, there are more inactive clones than active ones that consist purely of activators.

Reverse complementation can also affect function. Still considering only activators, the orientation ><< is active (L22-47), while >>< (L22K-102) is not. Here again, a small difference in fluorescence translates into survival. Also the particular random element that is present is material; *e.g.*, <4|A>|A> (L22K-87) and

TABLE 3
Statistics describing the active and inactive pools of artificial URS, mean \pm SD

Library L22 A = MSE-DIT	Length	Fluorescence	A>: random	<A : random	A>: (A> + <A)	AA> + <AA	A><A + <A A>
Before selection	4.92 \pm 1.91	0.00 \pm 0.01	1.07 \pm 0.89	0.87 \pm 1.12	0.56 \pm 0.32	0.93 \pm 0.95	0.84 \pm 0.84
After selection	6.09 \pm 2.15	0.18 \pm 0.06	1.34 \pm 1.32	1.39 \pm 1.21	0.49 \pm 0.25	1.04 \pm 1.22	1.37 \pm 1.30
Library L23 A = SSE-DIT	Length	Fluorescence	A>: random	<A : random	A>: (A> + <A)	AA> + <AA	A><A + <A A>
Before selection	5.75 \pm 2.58	0.00 \pm 0.01	1.04 \pm 1.26	0.72 \pm 0.89	0.54 \pm 0.32	1.00 \pm 1.26	0.67 \pm 0.81
After selection	7.35 \pm 2.71	0.32 \pm 0.18	1.24 \pm 1.09	1.26 \pm 1.09	0.50 \pm 0.25	1.48 \pm 1.37	1.31 \pm 1.31

Length is in units of number of elements. The next three columns are ratios of element numbers (clone skipped if denominator 0), and the last two columns, the number of activators in the configurations >> + << (direct) and <> + >< (indirect).

|A>>|4>>|A> (L22K-81) are inactive, while |3>>|A>>|A> (L22-100) and |A>><4|A> (L22-97) are active.

Similar trends are observed in the libraries built from the other activator SSE-DIT (supplemental Table S2 at <http://www.genetics.org/supplemental/>). There is overrepresentation of both shorter clones and those consisting purely of activator elements in the inactive library, *e.g.*, three of the four possible combinations of two activators do not express (<>, ><, <<), while there are no length 2 clones in L23 at all. However, an inactive pair of activator elements when supplemented with another activator, as in <A|A><A| (L23-76), or random elements (upstream of the gene) |1>|3><A|A> (L23-9) become functional. A similar contrast is provided by |A>>|A><A| (L23K-55) *vs.* |4>>|A>>|A><A| (L23-49). The random element that sits between a pair of activators matters, as in |A>><2|A> (L23K-70) *vs.* |A>><4|A> (L23-29). (The same combination, |A>><4|A> with MSE-DIT as A, was active in L22.)

In a majority of these examples, a factor of 2–3 in fluorescence was responsible for survival under two independent selections. Furthermore, the examples that we have highlighted make it implausible that there is a strict binary rule distinguishing active and inactive clones. The other possibility is that expression is a “quantitative trait” and depends in a graded way on many aspects of the sequence. This makes the space of potential rules much larger, and algorithms that search for explanatory rules in such systems tend to rely on probabilistic evidence. This puts a premium on large data sets, whereas with binary logic relatively small data sets such as ours can exclude any rule for which a single counter example can be found. In addition, our active libraries do not exhibit a large range in fluorescence; *e.g.*, only three clones in L22 do not lie in the range 0.1–0.2. Therefore, while the determinants of expression may be quantitative, we do not observe a large range of activities.

In Table 3 we enumerate various traits (the ratio of activators to random elements, their orientations, etc.) that could control expression and then enumerate the

number of instances in the various libraries. There is no single trait that could reliably predict survival of a clone since the difference in means is much less than the combined standard deviation. However, a sample of 100 clones of either type could be classified, so the traits have predictive value. Hence we did a linear fit of the fluorescence to the six traits in Table 3 for all 185 clones containing MSE-DIT (L22) and the 190 clones with SSE-DIT. These traits accounted for 15% and 12% of the variance in the data, which is still very significant (*F*-test probabilities of 3×10^{-4} and 1.4×10^{-3} , respectively). For comparison, the decrease in variance when fitting microarray expression data to sequence motifs is often in the range 10–25% (BUSSEMAKER *et al.* 2001). However, this comparison is biased since for the array data most genes do not respond, while half of our clones provide meaningful signal. For each activator, we also used the k-means clustering algorithm (<http://www.mathworks.com/products/statistics/>) to partition the merged pre- and postselection libraries into two clusters on the basis of traits in Table 3. One cluster with ~65% of the data had equal representation from the pre- and postselection libraries. The second cluster had a 2:1 or 3:1 excess of the postselection clones, depending on the activator. Multiple traits contributed to the second cluster and the postselection clones clustered therein did not display unusually high fluorescence.

Our data do not allow a meaningful investigation of phasing with respect to the DNA helix (random elements with sticky ends were length 24 and the two activators length 31).

There is a large branch of computer learning theory devoted to classification on the basis of training examples. To successfully generalize, the algorithm must take account of the complexity of the model to avoid overfitting. One application analogous to the situation that we face with the two promoter libraries is deciding whether a newspaper story is about a certain subject. The training sets are stories labeled as to subject. A set of features commonly used to represent such data are the words that it contains, with each word weighed by some

measure of how informative it is within the English language and by how often it occurs in the story. The problem is then to classify an unknown story. The classifier that we use [a so-called support vector machine (SVM) (JOACHIMS 2002) and <http://svmlight.joachims.org/>] represents the story as a vector in the space of all words (each word is a direction, and the weight defines the projection along that direction). It then determines the best hyperplane that separates the positive and negative training examples.

Whereas for English it is plausible that words are informative features, it is not obvious what to choose for promoters. We constructed four different dictionaries of "words." To emphasize longer words, yet keep their number restricted, we reduced each promoter to a string over a three-letter alphabet representing the two orientations of the activator element and any random element, irrespective of orientation. Dictionaries (i) and (ii) used all 2-mers or 3-mers in the reduced alphabet and counted their number in each promoter shifting by one letter each time. Each of the 9 or 27 words was then given a weight proportional to the number of its occurrences. Another class of dictionaries retained separate random elements. In case (iii), 6 features were counted (activator with orientation, random element either orientation) and in case (iv) 32 features (all pairs of elements with one activator with all orientations included). The last two models are sensitive to cryptic sites either internal to the random elements or on the boundaries between them and the activator.

We assayed the performance of the SVM by leaving out one training example, fitting to the remainder, and then predicting the omitted data. For all dictionaries the performance was not much better than random. The reasons for this failure probably lie with the choice of dictionaries. For the two larger dictionaries, (ii) and (iv), we found the word that best discriminated the active from the inactive clones. The probability of the observed frequency bias between the two sets was consistent with chance when the size of the dictionaries was taken into account. To further quantify the discriminatory power of these dictionaries, we took the six most biased words and did a linear regression fit to the fluorescence (merging the pre- and postselection clones separately for each activator). About 10% of the sample variance was fit this way, a bit less than was achieved with the features from Table 3. A decision tree (<http://www.mathworks.com/products/statistics/>) based on the counts of words in each promoter worked no better than linear regression, probably because there is no natural hierarchy among the words.

DISCUSSION

When we began this project our expectation was that random combinations of single activating elements and spacers in any orientation would be active, and we were

surprised to find that only a few percentages of such clones were active. In contrast, libraries prepared from activators only (in random orientations) gave ~50% functional clones. Thus, barring artifacts attributable to the choice of elements, a significant proportion of the expression potential of a promoter resides in the arrangement of sites relative to potentially neutral sequence, not merely their number and orientation. It is notable that natural promoters essentially never consist of multimers of potential binding sites without intervening sequence (although this is a common experimental strategy for measuring URS activity), while multiple sites separated by intervening sequences that appear to be neutral are highly common features of natural promoters.

We developed a screen on the basis of the sensitivity of yeast ascospores lacking *Dit1* to treatment with lytic enzymes. In comparison to a more conventional screen based, for example, on auxotrophic markers, our approach allowed us to study a system that reached a fixed final point with a readout, dihydrotyrosine, that integrated the activity of the gene product that we assayed. We thus minimized the impact of timing differences and eliminated one potential component of the phenotype. Another possible approach is to use fluorescence-activated cell sorting of cells expressing a fluorescent marker. Such methods have the advantage of screening directly for level of gene expression, but suffer from low signal/noise ratio and low throughput, which would not allow us to select clones with a sensitivity of $1:10^4$.

We used both the MSE-DIT and SSE-DIT putative transcription factor binding sites as models for our screen for active arrangements of sites in URS. The screen was highly selective and its consistency was confirmed by alternative testing methods. SSE-DIT was probably the stronger of the two activators, judging by the fraction of the activator-only libraries that survived selection and the overall dihydrotyrosine fluorescence level. However, there was not a meaningful difference between the numbers of activators in the active *vs.* inactive sets for the libraries constructed with inclusion of spacers along with the active elements.

Some of the clones classified as inactive by our selection methods contained mono-concatenates and short (two to three elements) poly-concatenates of the putative binding sites. This may at first sight seem to be at odds with the established method of testing transcription factor binding site activity by placing several copies of the site upstream of a reporter gene, for example (RAI *et al.* 1989). However, for historical reasons, in this and many other reports the reporter is usually β -galactosidase downstream of the *CYCI* minimal promoter. The *CYCI* promoter contains two TATA boxes, one of them being permanently occupied by TATA-binding protein even in the absence of URS (CHEN *et al.* 1994; KURAS and STRUHL 1999; LI *et al.* 1999). This raises the possibility that TATA-binding protein

cooperatively aids transcription factor access to URS sites; this effect was observed in the case of a *MEL1* promoter construct similar to the one that we used, but cooperativity was ~3.5-fold stronger in the case of the *CYC1* promoter (VASHEE and KODADEK 1995). This may be cooperatively mediated by nucleosomes or chromatin structure. We therefore hypothesize that our screening system discriminates more stringently against weak sites/configurations of sites (see also MELCHER *et al.* 2000).

In designing random libraries there is tension between working with too limited a repertoire of elements that introduce artifacts due to repetition and too rich a collection of elements that make the space of the possible combinations too large for sampling. We wanted to be sure of a reasonable sampling of possible promoter structures, so our input fragments were a low-complexity mix, but this has the disadvantage that the presence of exact and inverted repeats in our promoters could generate atypical secondary structures. Another trade-off is between fidelity to the endogenous context (*i.e.*, we used 27 bp containing the DIT1–MSE, not a 10-bp consensus sequence) and the introduction of additional complexity. There could also be new binding sites introduced at the boundaries between elements although we could not detect such sites computationally. The size of our random elements was not a multiple of the DNA helical repeat, mitigating against cooperative interactions, and neither Ndt80 or Msn2,4 is known to require a cofactor. Nucleosome positions may account for our results but are impossible to predict, as are specific interactions with the core promoter.

If the rules distinguishing the active from inactive library elements are really binary, then they must be very intricate, on the basis of changes in activity accompanying what would seem to be immaterial changes in sequence. What matters for the selection in the majority of cases is a quantitative change in di-tyrosine level, and at this refined level of readout the determinants of expression are multiple traits diffusively spread over the promoter. A similar conclusion was reached in SEKINGER *et al.* (2005). The selection appears to be quite sharp over a limited range in di-tyrosine levels. Whether this is an artifact or a property of many natural promoters we cannot say until more mutational screens of regulatory regions are done. If we cannot discern relevant rules within our restricted universe with 100 examples, will interspecies comparisons do better? There are very few genes with identical expression under all conditions, so having an entire genome may provide a larger collection of fruit, but not more depth of coverage.

In addition, we identified a potential new sporulation-specific regulatory site in the *DIT1/DIT2* intergenic region, SSE–DIT, active during the late spore-wall-forming phase of sporulation and inactive during vegetative growth. The predicted consensus sequence for this site, TRAGGGGY, is similar to the “canonical” stress response element in yeast AGGGG, which is bound

by transcription factors Msn2 and Msn4 (MARTINEZ-PASTOR *et al.* 1996), and possibly some paralogues such as Mig1 [binding consensus YGGGG (MUKHERJEE *et al.* 2004)]. The stress factors are upregulated early in sporulation (PRIMIG *et al.* 2000) so some additional effect would be needed to explain the late expression of our GFP reporter driven by the SSE–DIT-containing library. In the context of our libraries, SSE–DIT appears to be a stronger activator than MSE–DIT.

The regulatory region for *DIT1/DIT2* is large and surprisingly complex for genes that do not participate in regulation and are sporulation specific. In spite of considerable effort, conventional promoter dissection has not achieved a clear explanation of the time course of these genes. The available evidence suggests that the MSE–DIT mediates activation by Ndt80, on the basis of its similarity to the consensus, *in vitro*-binding assays, and the upregulation of *DIT1* in response to ectopic Ndt80 expression in vegetative cells. Ndt80 is not essential for *DIT1* activity during sporulation since there is no expression change in an *ndt80Δ* strain (CHU *et al.* 1998), and related species may not have a functional Ndt80 site. A plausible activator site is the SSE–DIT element, perhaps driven by Msn2,4. It is present at five to six places within the *DIT1/DIT2* regulatory region, most of which are conserved between species. Other activation (probably indirect) comes from Rim101 (BOGENGRUBER *et al.* 1998). However, if the stress factors do play a role, then it has to be explained why *DIT1* is not more consistently active in stress experiments; perhaps the repression is conveyed through the 76-bp NRE fragment (FRIESEN *et al.* 1997; BOGENGRUBER *et al.* 1998) as in vegetative growth or via Mig1 binding to SSE–DIT.

There is enrichment of the SSE–DIT site in the small set of 11 mid-late genes common to both array experiments, but not in any larger set that we could find. Because the SSE–DIT is much less widely represented among sporulation-specific genes than Ndt80 sites, it is a candidate for a site that may have evolved in response to the presence of an activating factor (*e.g.*, Msn2,4), followed by the decay of an ancient and now unnecessary Ndt80 site. The overlapping region where the repressor Sum1 (PIERCE *et al.* 2003; WANG *et al.* 2005) may bind is better conserved among *Saccharomyces* species.

Our results point to unexpected complexity and diffuseness in the rules governing the construction of functional promoters from activator sites. Complementary assays under other conditions for other activators are clearly needed, particularly if one library could be tested under multiple conditions (*e.g.*, our SSE–DIT library in a stress experiment). It will also be interesting if the behavior of inactive and active library clones is modified in strains lacking components of the chromatin-remodeling pathways.

We thank J. Widom for helpful discussions. This work was supported by the Burroughs Wellcome Fund (F.R.C., E.D.S.) and the National Science Foundation DMR–129848 (E.D.S.).

LITERATURE CITED

- BEAN, J. M., E. D. SIGGIA and F. R. CROSS, 2005 High functional overlap between MluI cell-cycle box binding factor and Swi4/6 cell-cycle box binding factor in the G1/S transcriptional program in *Saccharomyces cerevisiae*. *Genetics* **171**: 49–61.
- BOGENGRUBER, E., T. EICHBERGER, P. BRIZA, I. W. DAWES, M. BREITENBACH *et al.*, 1998 Sporulation-specific expression of the yeast DIT1/DIT2 promoter is controlled by a newly identified repressor element and the short form of Rim101p. *Eur. J. Biochem.* **258**: 430–436.
- BRIZA, P., G. WINKLER, H. KALCHHAUSER and M. BREITENBACH, 1986 Dityrosine is a prominent component of the yeast ascospore wall. A proof of its structure. *J. Biol. Chem.* **261**: 4288–4294.
- BRIZA, P., M. BREITENBACH, A. ELLINGER and J. SEGALL, 1990 Isolation of two developmentally regulated genes involved in spore wall maturation in *Saccharomyces cerevisiae*. *Genes Dev.* **4**: 1775–1789.
- BRIZA, P., M. ECKERSTORFER and M. BREITENBACH, 1994 The sporulation-specific enzymes encoded by the DIT1 and DIT2 genes catalyze a two-step reaction leading to a soluble LL-dityrosine-containing precursor of the yeast spore wall. *Proc. Natl. Acad. Sci. USA* **91**: 4524–4528.
- BUSSEMAKER, H. J., H. LI and E. D. SIGGIA, 2001 Regulatory element detection using correlation with expression. *Nat. Genet.* **27**: 167–171.
- CHEN, J., M. DING and D. S. PEDERSON, 1994 Binding of TFIID to the CYC1 TATA boxes in yeast occurs independently of upstream activating sequences. *Proc. Natl. Acad. Sci. USA* **91**: 11909–11913.
- CHU, S., J. DERISI, M. EISEN, J. MULHOLLAND, D. BOTSTEIN *et al.*, 1998 The transcriptional program of sporulation in budding yeast. *Science* **282**: 699–705.
- ESPOSITO, R. E., and S. KLAPHOLZ, 1981 Meiosis and ascospore development, pp. 211–287 in *The Molecular Biology of the Yeast Saccharomyces cerevisiae: Life Cycle and Inheritance*, edited by J. N. STRATHERN, E. W. JONES and J. R. BROACH. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- FRIESEN, H., S. R. HEPWORTH and J. SEGALL, 1997 An Ssn6-Tup1-dependent negative regulatory element controls sporulation-specific expression of DIT1 and DIT2 in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **17**: 123–134.
- HEPWORTH, S. R., H. FRIESEN and J. SEGALL, 1998 NDT80 and the meiotic recombination checkpoint regulate expression of middle sporulation-specific genes in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **18**: 5750–5761.
- IYER, V. R., C. E. HORAK, C. S. SCAFE, D. BOTSTEIN, M. SNYDER *et al.*, 2001 Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**: 533–538.
- JOACHIMS, T., 2002 *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Springer-Verlag, Berlin/Heidelberg, Germany/New York.
- JOLLY, E., C. S. CHIN, I. HERSKOWITZ and H. LI, 2005 Genome-wide identification of the regulatory targets of a transcription factor using biochemical characterization and computational genomic analysis. *BMC Bioinformatics* **6**: 275.
- KASSIR, Y., N. ADIR, E. BOGER-NADJAR, N. G. RAVIV, I. RUBIN-BEJERANO *et al.*, 2003 Transcriptional regulation of meiosis in budding yeast. *Int. Rev. Cytol.* **224**: 111–171.
- KURAS, L., and K. STRUHL, 1999 Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme. *Nature* **399**: 609–613.
- LAMOUREUX, J. S., D. STUART, R. TSANG, C. WU and J. N. GLOVER, 2002 Structure of the sporulation-specific transcription factor Ndt80 bound to DNA. *EMBO J.* **21**: 5721–5732.
- LI, X. Y., A. VIRBASUS, X. ZHU and M. R. GREEN, 1999 Enhancement of TBP binding by activators and general transcription factors. *Nature* **399**: 605–609.
- MARTINEZ-PASTOR, M. T., G. MARCHLER, C. SCHULLER, A. MARCHLER-BAUER, H. RUIS *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.
- MELCHER, K., B. SHARMA, W. V. DING and M. NOLDEN, 2000 Zero background yeast reporter plasmids. *Gene* **247**: 53–61.
- MILLER, J. A., and J. WIDOM, 2003 Collaborative competition mechanism for gene activation in vivo. *Mol. Cell. Biol.* **23**: 1623–1632.
- MITCHELL, A. P., 1994 Control of meiotic gene expression in *Saccharomyces cerevisiae*. *Microbiol. Rev.* **58**: 56–70.
- MUKHERJEE, S., M. F. BERGER, G. JONA, X. S. WANG, D. MUZZEY *et al.*, 2004 Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* **36**: 1331–1339.
- PIERCE, M., K. R. BENJAMIN, S. P. MONTANO, M. M. GEORGIADIS, E. WINTER *et al.*, 2003 Sum1 and Ndt80 proteins compete for binding to middle sporulation element sequences that control meiotic gene expression. *Mol. Cell. Biol.* **23**: 4814–4825.
- PRIMIG, M., R. M. WILLIAMS, E. A. WINZELER, G. G. TEVZADZE, A. R. CONWAY *et al.*, 2000 The core meiotic transcriptome in budding yeasts. *Nat. Genet.* **26**: 415–423.
- RAI, R., F. S. GENBAUFFE, R. A. SUMRADA and T. G. COOPER, 1989 Identification of sequences responsible for transcriptional activation of the allantoinase permease gene in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **9**: 602–608.
- REN, B., F. ROBERT, J. J. WYRICK, O. APARICIO, E. G. JENNINGS *et al.*, 2000 Genome-wide location and function of DNA binding proteins. *Science* **290**: 2306–2309.
- SEKINGER, E. A., Z. MOQTADERI and K. STRUHL, 2005 Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol. Cell* **18**: 735–748.
- SHERMAN, F., 1991 Getting started with yeast. *Methods Enzymol.* **194**: 3–21.
- SIGGIA, E. D., 2005 Computational methods for transcriptional regulation. *Curr. Opin. Genet. Dev.* **15**: 214–221.
- SIMON, I., J. BARNETT, N. HANNETT, C. T. HARBISON, N. J. RINALDI *et al.*, 2001 Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**: 697–708.
- VASHEE, S., and T. KODADEK, 1995 The activation domain of GAL4 protein mediates cooperative promoter binding with general transcription factors in vivo. *Proc. Natl. Acad. Sci. USA* **92**: 10683–10687.
- WANG, W., J. M. CHERRY, Y. NOCHOMOVITZ, E. JOLLY, D. BOTSTEIN *et al.*, 2005 Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc. Natl. Acad. Sci. USA* **102**: 1998–2003.

Communicating editor: A. P. MITCHELL