nature biotechnology

Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters

Eilon Sharon[1,2,5], Yael Kalma[1,2,5], Ayala Sharp[2], Tali Raveh-Sadka[1], Michal Levo[1], Danny Zeevi[1,2], Leeat Keren[1,2], Zohar Yakhini[3,4], Adina Weinberger[1,2] & Eran Segal[1,2]

Despite extensive research, our understanding of the rules according to which *cis*-regulatory sequences are converted into gene expression is limited. We devised a method for obtaining parallel, highly accurate gene expression measurements from thousands of designed promoters and applied it to measure the effect of systematic changes in the location, number, orientation, affinity and organization of transcription-factor binding sites and nucleosome-disfavoring sequences. Our analyses reveal a clear relationship between expression and binding-site multiplicity, as well as dependencies of expression on the distance between transcription-factor binding sites and gene starts which are transcription-factor specific, including a striking ~10-bp periodic relationship between gene expression and binding-site location. We show how this approach can measure transcription-factor sequence specificities and the sensitivity of transcription-factor sites to the surrounding sequence context, and compare the activity of 75 yeast transcription factors. Our method can be used to study both *cis* and *trans* effects of genotype on transcriptional, post-transcriptional and translational control.

Deciphering how motifs present in DNA sequences are linked to gene expression is crucial to understanding how transcriptional regulation works. The quantitative effect on gene expression of even the most basic organizational features of promoters is not well understood. For example, even for a single transcription-factor binding site, we know little about the quantitative effects of its location, orientation and affinity on gene expression, whether these effects are general, factor specific and/or promoter dependent and how they depend on the underlying nucleosome organization.

In principle, such issues can be resolved through accurate measurements of promoter expression in which transcription-factor binding motifs are systematically varied. Several medium-scale[1–3] and large-scale[4–6] libraries have been created in bacteria and yeast in which transcription regulatory elements were randomly ligated or mutagenized and the expression of the resulting promoters was measured. Although these studies have provided useful insights into the control of transcription, they are not well suited, for example, to address the effect of binding-site location on gene expression. This would require analyses of multiple promoters that differ only in the location of the transcription-factor binding site. Randomly ligated libraries of promoter regulatory elements would not contain all possible combinations of different locations of transcription-factor binding sites. The measurement of gene expression from specifically designed promoter variants[7–10] has led to profound insights, such as the detection of cooperativity in DNA binding of transcription factors[7], but because the promoter variants were constructed one at a time, both time and cost considerations have limited the scale of previous studies of promoter organization to, at most, dozens of variants.

One recent study showed the benefit of using thousands of designed sequences for analyzing the effect of systematic mutations of three bacteriophage promoters and three mammalian core promoters[11]. However, this method measures promoter strength *in vitro* and therefore has limited utility for understanding promoter activity *in vivo*. While our paper was in review, two recently published methods have used parallel measurement of promoter activity *in vivo*[12,13]. One of these methods assayed the effect of a library of >100,000 random mutations in three mammalian enhancers on gene expression[12], but the random nature of the library limits the use of this method for the systematic dissection of regulatory logic. The other method[13] used programmable microarrays[14] to measure the effect of systematically designed mutations in two mammalian enhancers.

In this study we devised a high-throughput fluorescence-based method to obtain parallel and highly accurate expression measurements of thousands of fully designed promoters in yeast. Our approach differs from those of the two studies mentioned above[12,13] in several ways. First, the parallel expression measurements we obtained are in better agreement with those of isolated strains ($R^2 = 0.99$) than the measurements reported by one of the previous studies[13] ($R^2 = 0.45$–0.75). Highly accurate expression measurements are necessary for a quantitative understanding of transcriptional regulation. Second, in

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [3]Computer Science department, Technion, Haifa, Israel. [4]Agilent Laboratories, Santa Clara, California, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to E. Segal (eran.segal@weizmann.ac.il) or A.W. (adina.weinberger@weizmann.ac.il).

Received 5 March; accepted 4 April; published online 20 May 2012; doi:10.1038/nbt.2205

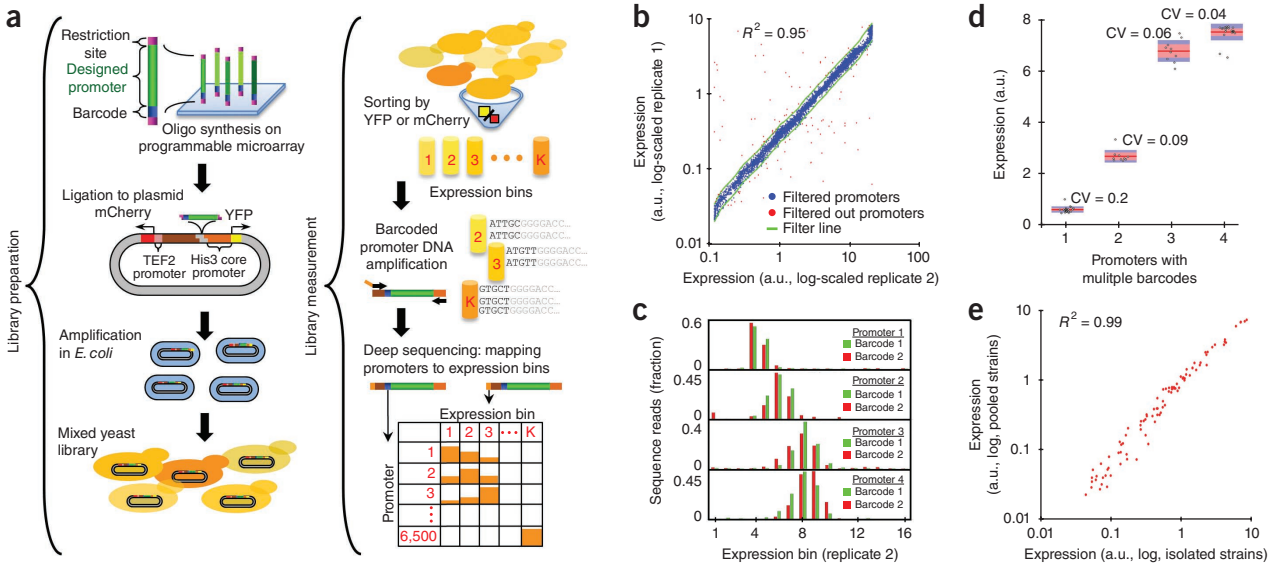

Figure 1 Obtaining accurate expression measurements for thousands of designed promoter sequences. (**a**) Illustration of the experimental method. (**b**) Comparison of the expression measurements (log scale) obtained for two independent replicates done using two different cell-sorting strategies. Of the 6,500 promoters that we designed, 114 (1.75%) fell outside the green lines (30% difference from mean of two replicates) and were filtered out of the analyses. A.u., arbitrary units. (**c**) Barcodes have little effect on expression measurements. Distribution of sequencing reads across the expression bins obtained for four pairs of promoters that differ only in their barcode sequence. (**d**) As **c**, but distribution of reads for four sets of promoters where each set contains 10 (columns 3 and 4) or 20 (columns 1 and 2) promoters that differ only in their barcode sequence. Individual expression measurements, gray dots; their median, red lines; standard error, pink bars; s.d., blue bars; CV (s.d. divided by the mean). (**e**) A comparison of the expression for 92 individual strains isolated from the pool of transformed yeast cells and sequenced. Each strain was measured in isolation using a flow cytometer (*x* axis) or within a single experiment using our method (*y* axis).

contrast to the methods used in the two previous studies[12,13], our method does not require the use of barcodes to tag the RNA reporter, as each designed promoter (and not the mRNA as in refs. 12,13) is sequenced. Barcodes within the RNA can affect reporter expression and limit the accuracy of gene expression measurements[13]. It should be noted, however, that our study incorporated a barcode upstream of the designed promoter to enable the use of short (40-bp) sequencing reads, this barcode can be avoided by fully sequencing the designed promoter (103 bp). Third, although both previously published methods measure the mean expression over a cell population, our method obtains cell-to-cell (noise) expression variability measurements for each promoter, which also agree well with the isolated strain measurements (R^2 = 0.43; **Supplementary Fig. 1**). Fourth, by using protein fluorescence and not RNA as the readout of gene expression, we can also study translational control, for example, using libraries that alter the 5′ untranslated region or the codons of the fluorescent reporter. In addition, the need to physically couple a proximal barcode to the examined variable region limits both of the previously described methods[12,13] to studying *cis* effects, whereas our method can be used to examine the effects of sequence variation on fluorescent protein expression in *trans*. We note, however, that RNA-based measurements are advantageous for certain applications, such as studying fast transcriptional dynamics.

We designed a library of 6,500 yeast promoters that directly measures several grammatical rules of transcriptional regulation, such as the effect on gene expression of binding-site location, number, orientation and affinity. Our results include identification of a clear logistic-function relationship between gene expression and transcription-factor binding-site number; a dominance of transcription-factor identity over site number in determining high expression; a large effect on gene expression of even small (1–7 bp) changes in transcription-factor binding-site location; and, for one transcription factor, a striking ~10-bp periodic relationship between gene expression and transcription-factor binding-site location. Our approach can be adapted to other genomic regions and organisms to unravel diverse types of *cis* and *trans* relationships between sequence and phenotype.

RESULTS

Pooled expression measurements of designed promoters

We designed a library of 6,500 different yeast promoters that address diverse questions in transcriptional regulation and devised a method for accurately measuring their expression in a single experiment (**Fig. 1a**). Briefly, we obtained a mixed barcoded oligonucleotide pool synthesized on Agilent programmable microarrays[11,14,15] that represents our promoter library and inserted it into a low-copy plasmid such that it was inserted upstream of a ~100-bp TATA-containing core promoter followed by a yellow fluorescent protein (YFP) reporter gene. We then amplified the library in *Escherichia coli* and transformed it into yeast. We next sorted the resulting pool of transformed cells, grown in a desired condition, based on YFP intensity (**Fig. 1a**) and used deep sequencing to obtain a measure of the expression of each promoter based on the distribution of its sequencing reads across the sorted gene expression bins.

We designed a large fraction (55%) of our library using binding sites for the two well-studied transcriptional activators Gal4 and Gcn4. Accordingly, we grew the cells in galactose medium while starving them of all amino acids except histidine and leucine, as this condition activates both transcription factors. To test the generality of our conclusions, we performed all of the systematic changes to the regulatory elements in two different promoter backgrounds: “mGal1-10” background which is a sequence from the native yeast Gal1-10 promoter and “mHis3” background which is a sequence from the yeast His3 promoter. In both backgrounds the native regulatory elements were either mutated or deleted (as described in the online methods).

We used several tests to gauge the accuracy of our approach. First, all of the designed promoters were represented in the final sequencing reads, and 94% of the promoters had at least 100 reads. Second, we showed that our method is reproducible, as independent replicates sorted using two different strategies (three recursive sorting to four bins (replicate 1) and sorting directly to 16 bins (replicate 2)) were highly correlated ($R^2 = 0.95$; **Fig. 1b**). Third, we verified that the barcode had little effect on the reporter gene expression by designing 22 promoters, each with 2-20 different barcodes, and finding good agreement between the gene expression measured from promoters which differed only in their barcode (**Fig. 1c**,**d** and **Supplementary Fig. 2**). Fourth, and most critically, we isolated 92 individual clones from the mixed pool of transformed yeast cells, sequenced each of them to identify the integrated promoter and measured the expression of each isolated clone individually using flow cytometry. Notably, we found excellent agreement ($R^2 = 0.99$; **Fig. 1e**) between these measurements and measurements of the same promoters using our high-throughput method. Fifth, because our promoters are on plasmids, we compared the expression of 29 of our promoters to the measurements of the expression of 29 corresponding individual strains in which the full DNA sequence of our construct including the designed promoter and the reporter genes were integrated into the genome at a fixed location and again found excellent agreement between the gene expression under the control of the plasmid-borne and the chromosomally integrated designed promoters ($R^2 = 0.97$; **Supplementary Fig. 3**) suggesting that our plasmid-based assay reflects the function of chromosomally located promoters.

Together, these results show that this method can measure the expression of a reporter gene under the control of thousands of fully designed promoters in a single experiment with similar accuracy to that obtained when the promoters are constructed and assayed individually.

Scanning mutagenesis detects functional promoter elements

We first examined the utility of our method to comprehensively map functional promoter elements. We selected 103-bp regions from three native yeast promoters and designed individual systematic mutations across all of their nonoverlapping 4-bp segments. This scanning mutagenesis method can identify regulatory elements[11,16], and we found a substantial reduction in expression when putative transcription-factor sites were mutated (**Supplementary Fig. 4**). Notably, we found similarly strong reductions in reporter gene expression when a poly(dA:dT) tract, which disfavors nucleosome formation[17–19] (**Supplementary Fig. 4c**), was mutated, suggesting a previously unrecognized regulatory role for this region. In contrast, mutation of two putative transcription-factor sites in a different promoter had little effect (**Supplementary Fig. 4b**), which might indicate that these sites are not functional in our assay conditions. Because we can measure thousands of promoters at once, these results confirm that our method can systematically map functional regulatory elements in promoters.

Profiling the activity of yeast transcription factors

We next used our method to compare the activity of 75 different yeast transcription factors by separately inserting their published consensus sites[20] into the same promoter in two possible orientations. This set of designed promoters compares transcription-factor activity by the expression that the consensus site of each induces in the same promoter context and growth condition and provides an alternative to comparisons based on protein abundance[21] and cellular localization[22], which do not capture the dependence of transcription-factor activity on parameters such as post-translational modification state and cofactor activity.

Of the transcription-factor sites that we tested, the expression of 53% was comparable to a null promoter with no site, suggesting that, at least in the conditions used in these experiments, these sites have little effect on gene expression on their own (**Fig. 2a**). Of the remaining sites, 24% and 23% had higher and lower expression than the null promoter, respectively, and their cognate transcription factors corresponded to known activators (e.g., Rap1 (ref. 23) or Aft2 (ref. 24)) and repressors (e.g., Rim101 (ref. 25) or Cin5 (ref. 26)), respectively, validating our assay for profiling transcription-factor activity. Notably, these data provided the first direct test of the *in vivo* activity of some of the consensus transcription-factor binding sites, thereby suggesting previously unidentified regulatory roles for their cognate transcription factors. For example, the consensus binding site for Ecm23, which our results designated as repressing gene expression, was previously reported as a repressor of pseudohyphal growth[27]. Deletion of the gene *YER184C*, whose site we identified as activating gene expression, prevents growth on glycerol or lactate as a sole carbon source[28] (**Fig. 2a**). Neither of these transcription-factor binding site activities had previously been experimentally validated. In addition, by comparing the expressions of the two different orientations of each transcription-factor binding site, we obtained a measure of the site-orientation effect and found significant effects for only 6 (8%) of the 75 tested transcription factors ($P < 0.05$, with effects ranging from1.9-fold to 2.3-fold; **Supplementary Fig. 5**). Among these six transcription factors was Rap1, which is consistent with results from a prior mutational analysis[29] and with an orientation bias for its sites in Rap1 target promoters[30].

Taken together, although these results may depend on the tested promoter context or growth condition, they directly compare the activity of many transcription-factor sites, suggest new regulatory roles for several transcription factors and quantify the transcriptional effect of transcription-factor binding-site orientation.

The effect of binding-site affinity

Despite its key role, systematic assays of the effect of transcription-factor binding-site affinity on expression are not available. We propose that our method can be used to perform such assays by comparing the expression of promoters in which only the transcription-factor binding site is systematically varied. To show this, we separately inserted the consensus sites of three different transcription factors into the same promoter background, along with all possible single–base-pair mutations of that consensus and many mutations comprising combinations of two and three base pairs. For Gcn4, the expression of both the consensus and its reverse complement were more than threefold higher than all other site variants, which themselves generated a continuous range of expressions (**Supplementary Fig. 6a**). Notably, we found good agreement ($R^2 = 0.93$; **Supplementary Fig. 6b**) between these expression measurements and those predicted by the *in vitro* Gcn4 site affinities[31], which persisted even at the lower expressions and affinities, suggesting that even for weak sites, affinity differences are manifested *in vivo*. Sites for the two other transcription factors, Fhl1 and Leu3, had lower expressions than Gcn4, and their measurements were thus less accurate. Nevertheless, these data also showed good correlation to *in vitro* measurements ($R = 0.21–0.28$), and for Fhl1, measurements of gene expression from mutagenized promoters provided the first comprehensive *in vivo* validation of the *in vitro* binding specificities previously identified for this transcription factor[20] (**Supplementary Fig. 7**). These results support the use of our

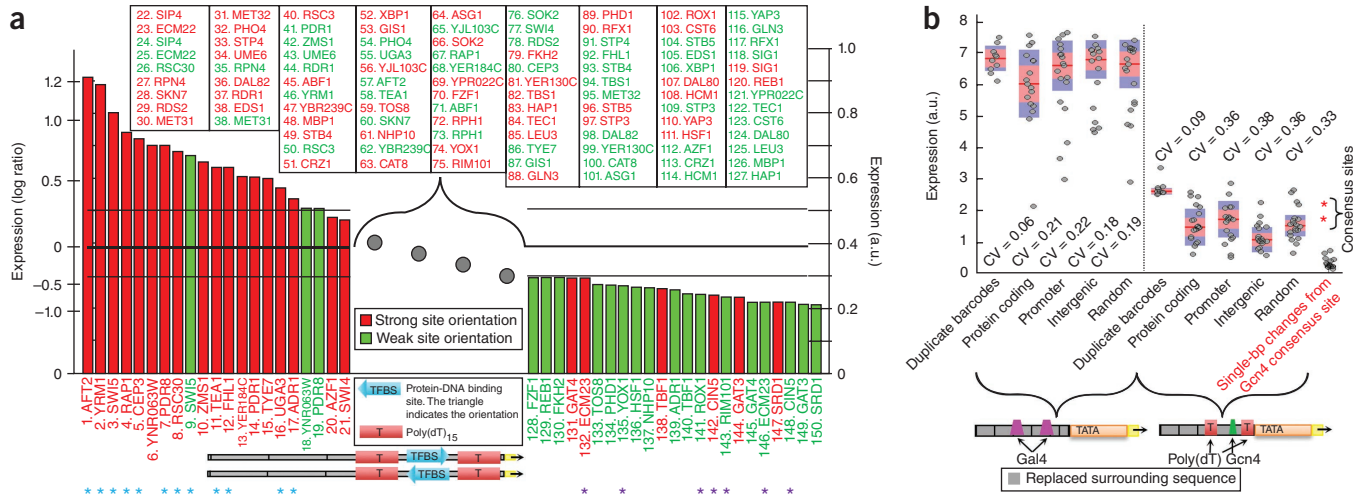

Figure 2 Profiling the activity of 75 yeast transcription factors. (**a**) Consensus binding sites for factors were separately inserted in both possible orientations at the same position in a fixed promoter context (bottom). Shown is a ranking of the resulting expression of each promoter, with the two site orientations of each transcription factor colored red and green depending on whether they correspond to the orientation with higher or lower expression, respectively. Individual measurements for promoters with intermediate expressions are not shown (transcription-factor sites and their internal rankings are indicated in the box). Blue and purple stars mark transcription factors with literature-reported activating or repressing roles, respectively. Thick horizontal black line, expression of the same fixed promoter described above but without any known transcription-factor binding site; two thin lines, confidence level of 30%. (**b**) The surrounding sequence has a significant but limited effect on the expression of regulatory elements and is similar for different types of surrounding sequences. Shown are the expressions of promoters in which a regulatory block consisting of two Gal4 binding sites (left five columns) or a single Gcn4 binding site flanked by two nucleosome-disfavoring sequences (right five columns) was placed at the same position within different types of surrounding sequence contexts. The sequence contexts were chosen randomly from yeast protein-coding regions (20 sequences), yeast promoters (20 sequences) or yeast intergenic regions that are not promoters (20 sequences), and 20 sequences were generated randomly using the same G/C content as that of yeast promoters (G/C = 40%, 20 sequences). For comparison, each regulatory block was also placed 20 different times within the same promoter, each time with a different barcode (columns 1 and 6). For each set, the individual promoter expression (gray dots) and the median (red line), standard error (pink bars), s.d. (blue bars) and CV (s.d. divided by the mean) are shown. As another comparison for the effect of the surrounding sequence on expression, the rightmost column shows the expressions of all 21 promoters from **Supplementary Figure 6a** in which a single base pair in the Gcn4 consensus site (gray points) was mutated, along with the expression of a promoter that contains either the consensus or its reverse complement (red stars).

method for assaying the effect of site affinity *in vivo* and suggest that *in vitro* site-affinity assays[31–33] provide a reliable measure of the effect of binding-site sequence across a broad range of affinities.

The effect of sequence context on regulatory motifs

We next tested the effect of varying the promoter background on the reporter gene expression that is controlled by two blocks of regulatory elements, one consisting of two Gal4 sites and another of a single Gcn4 site flanked by two poly(dA:dT) tracts. We separately embedded each block at a fixed position within 80 different surrounding sequences, which we selected randomly from yeast protein-coding regions (20 sequences), yeast promoters (20) and nonpromoter intergenic yeast regions (20), and a further 20 sequences generated randomly using the ~40% G/C content of native yeast promoters. The expression variability of each set of 20 promoters (coefficient of variation (CV) = 0.18–0.38) was greater than the variability obtained when placing these same regulatory blocks in 20 promoters that differed only by barcode (CV = 0.06–0.09; **Fig. 2b**). However, although significant ($P < 0.03$ in seven of eight cases), these context effects were smaller than the effect of single–base-pair mutations in the transcription-factor binding site, and nearly all (99%) of the promoters with two Gal4 sites induced markedly higher reporter gene expression than did all promoters with a single Gcn4 site (**Fig. 2b**). Notably, for both regulatory blocks, the distribution of expression was similar between the four different sequence contexts tested. Together, these results suggest that sequences that surround regulatory elements can have substantial effects on expression, but the identity of the transcription-factor sites may be a stronger determinant of the resulting expression.

The effect of binding-site location

We systematically tested the effect of binding-site location on expression. We selected three transcription factors and separately inserted their consensus sites into 16 different promoter contexts, varying the site location in each context by increments of 1–4 bp. For 14 additional transcription factors, we designed similar constructs but varied the site location by 7-bp increments. Notably, for most transcription factors and contexts, the reporter gene expression and site location were related by a jagged function that was specific to the combination of the transcription-factor site and context, such that even small 1- to 7-bp changes in site location had major effects (**Fig. 3** and **Supplementary Figs. 8** and **9**). These effects are only partly explained by noise in our experiment (**Supplementary Fig. 10**), the presence of promoter barcodes (**Supplementary Fig. 11**), the removal of sequences in the original promoter that were replaced when transcription-factor sites were inserted (**Supplementary Fig. 12**) or the base pairs that flanked the inserted sites.

Beyond these jagged relationships, we found an overall trend of lower reporter gene expression, on average, as activator sites are placed farther away from the start of the gene, and we found an opposite trend for repressor sites (**Fig. 3c**,**d** and **Supplementary Fig. 9**). We did not find a clear trend in the effect of the repressor site when its location was held fixed and the location of an activator site was

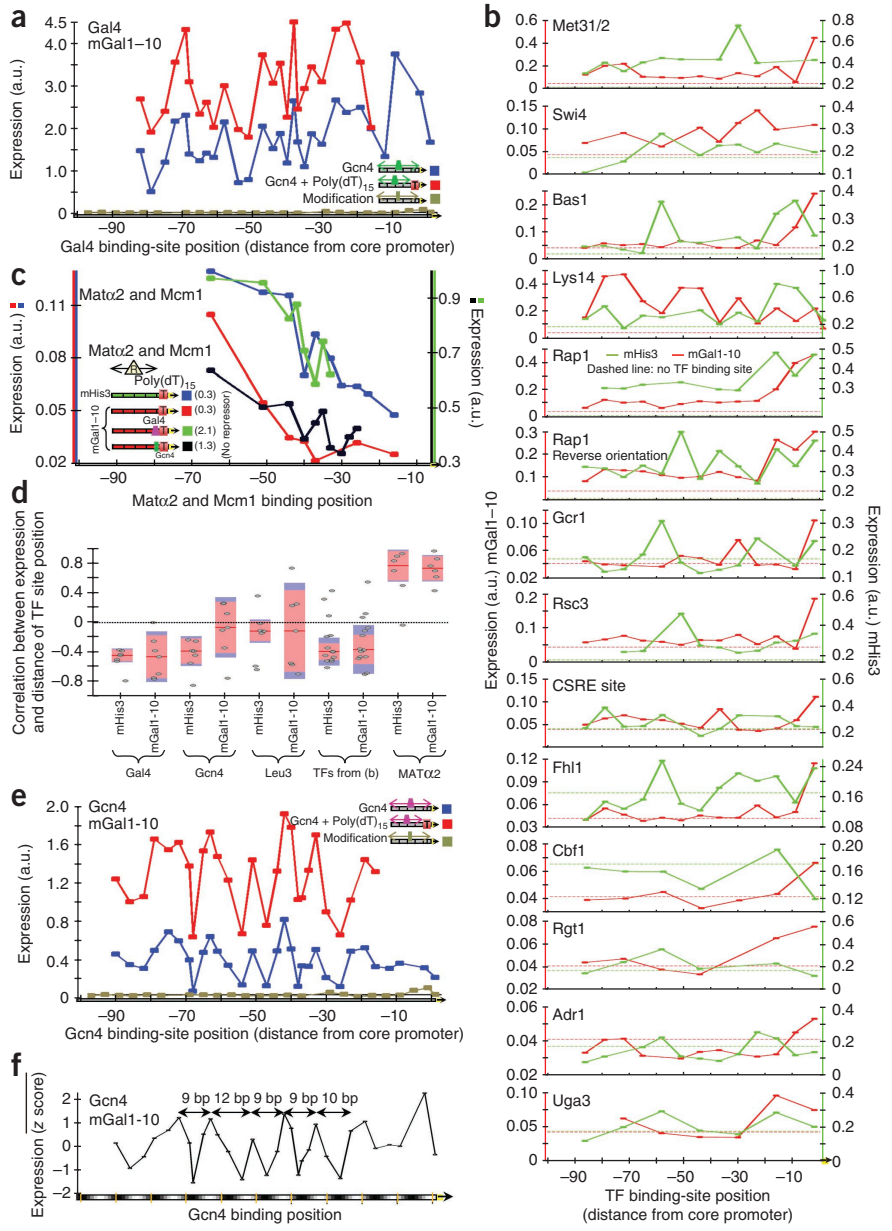

Figure 3 The effect of binding-site location on expression. (**a**) Expression of promoters depends on Gal4 site location. Points correspond to the location in the promoter of the rightmost base pair of the Gal4 site. For comparison, also shown are the expressions of the original promoter with no Gal4 sites (black line) and promoters (gray) in which random 3-bp mutations were performed across the non-poly(dA:dT) promoter, indicating that the effect of changing the location of the Gal4 sites is not a result of the removal of the original promoter sequence. (**b**) As in **a**, but shown are the expressions of 14 additional transcription factors whose sites we varied at 7-bp increments in two different promoter backgrounds. (**c**) The effect of the repressor sites decays with their distance from the core promoter. For the Matα2p-Mcm1p repressor complex, shown are four sets of promoters in which we modified the location of its repressor site along the promoter, where the four sets differ by the presence of poly(dA:dT) tracts and sites for the transcriptional activators Gcn4 and Gal4. For each of the four sets, the expression of the promoter without the repressor site is indicated in the inset legend and is higher than all promoters that contain the repressor site. (**d**) The effect of transcription factors on expression shows a general trend of decay with the distance between their sites and the core promoter. For each set of promoters in which we changed the location of a transcription-factor binding site within the same promoter background, we computed the correlation between the expression at each location and the distance of the transcription-factor site at that location from the core promoter. The resulting correlations are shown, where for Gal4, Gcn4, Leu3 and Matα2p-Mcm1p, each column groups together correlations of promoter sets for the same transcription factor in backgrounds that differ in the presence of poly(dA:dT) tracts, and for all other transcription factors that were each done in two distinct promoter backgrounds, correlations are grouped by backgrounds. For each column, the median (red line), standard error (pink bars) and s.d. (blue bars) of the correlations are shown. There is a trend of negative correlation between expression and site distance for all transcription factors except the repressor Matα2p-Mcm1p, for which there is a positive correlation. (**e**) Expression changes as a ~10-bp periodic function of Gcn4 site location. As in **a**, but for Gcn4 sites. (**f**) As in **e**, but here each point corresponds to the average expression of eight sets of promoters in which we changed the location of the Gcn4 site, where the eight different sets differ in the location of a poly(dA:dT) tract of length 15 bp. To normalize the expression across the eight different sets, the expression is shown as a robust *z* score, calculated by subtracting the median and dividing by the s.d. of expression differences from the median. The ~10-bp periodicity of expression observed over five periods can be seen (distances between neighboring peaks of expression are indicated, with the *x* axis gray-scale colors matching a 10.5-bp periodicity).

changed (**Supplementary Fig. 13**). Notably, for Gcn4, one of the three transcription factors whose sites we varied at increments of 1–4 bp, expression and site location were related by a periodic function that persisted over six consecutive peaks and whose period was ~10 bp, roughly matching the DNA helical repeat (**Fig. 3e**). This periodicity was significant ($P < 0.005$) in only one of the two promoter backgrounds in which we varied the Gcn4 site locations, but in this background, we observed it in seven different background variants (**Fig. 3f** and **Supplementary Fig. 14**). To test whether this finding can improve our ability to predict gene expression from DNA sequence alone, we extended a thermodynamic model for transcriptional regulation to include an interaction energy term between Gcn4 and RNA polymerase II that depends on the helical phase and found that this model indeed improves the expression predictions of test promoters that were held out when fitting the model parameters (**Supplementary Fig. 15**).

We note that even if similar periodicities exist for the other 14 transcription factors that we tested, the 7-bp site-location increments that we designed for the binding sites of these transcription factors prohibit their detection.

Taken together, our results show a dependency of expression on transcription-factor site location, such that even small 1- to 7-bp differences can have major effects. Although expression and site location are related by a jagged function that is specific to the

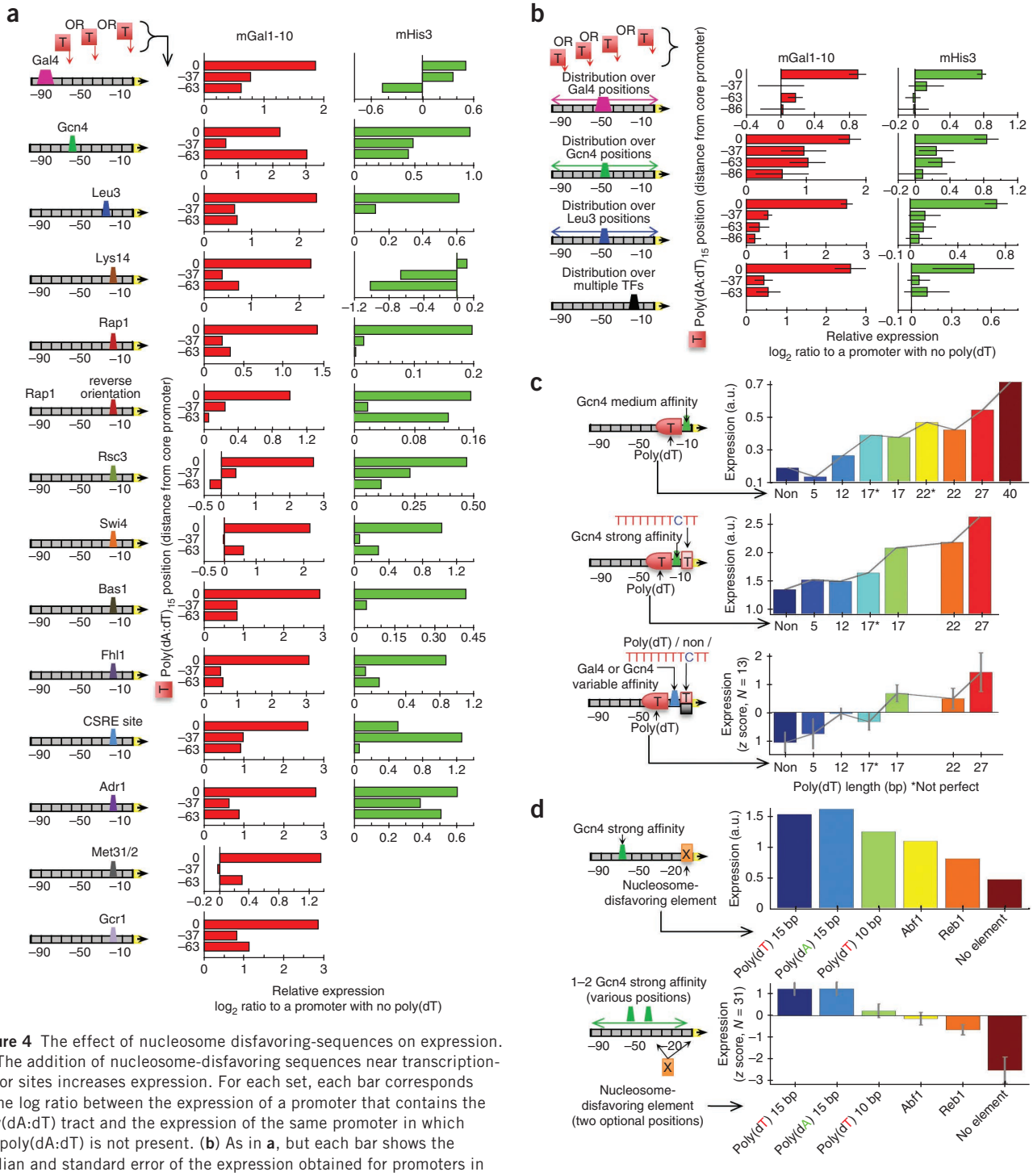

Figure 4 The effect of nucleosome disfavoring-sequences on expression. (**a**) The addition of nucleosome-disfavoring sequences near transcription-factor sites increases expression. For each set, each bar corresponds to the log ratio between the expression of a promoter that contains the poly(dA:dT) tract and the expression of the same promoter in which the poly(dA:dT) is not present. (**b**) As in **a**, but each bar shows the median and standard error of the expression obtained for promoters in which the poly(dA:dT) was at a fixed position and the location of the transcription-factor site varied. The fourth row ('multiple TFs') represents the average of the bottom 11 transcription factors shown in **a**. (**c**) The stimulatory effect of poly(dA:dT) tracts increases with their length. Shown are expressions for two sets of promoters (first two rows) in which sites with different affinities for Gcn4 were separately placed at a fixed location within different promoter backgrounds that contained poly(dA:dT) tracts of varying lengths at a fixed promoter location ('Non' represents a promoter with no poly(dA:dT)). Also shown (bottom row) are the median and standard error of the expression for promoters with various transcription-factor sites and site affinities. (**d**) The stimulatory effect of poly(dA:dT) tracts can be greater than that of the general transcription-factor activators Reb1p and Abf1p. Shown are the expressions of promoters in which different elements (no element, Reb1p site, Abf1p site, 10-bp poly(dA:dT) tract, 15-bp poly(dA:dT) tract or 15-bp poly(dA:dT) tract with a flipped orientation) were placed at the same location within a promoter background that contains a consensus Gcn4 site at a fixed location (top row). For each element, also shown are the mean and s.d. of the expression of promoters in which the element was inserted at two possible positions within 31 different promoter backgrounds that differ in the number and location of Gcn4 sites and the surrounding sequence (bottom row). To normalize the expression across the promoters of each set, expression is shown as a robust *z* score, calculated by subtracting the median and dividing by the s.d. of distances from the median.

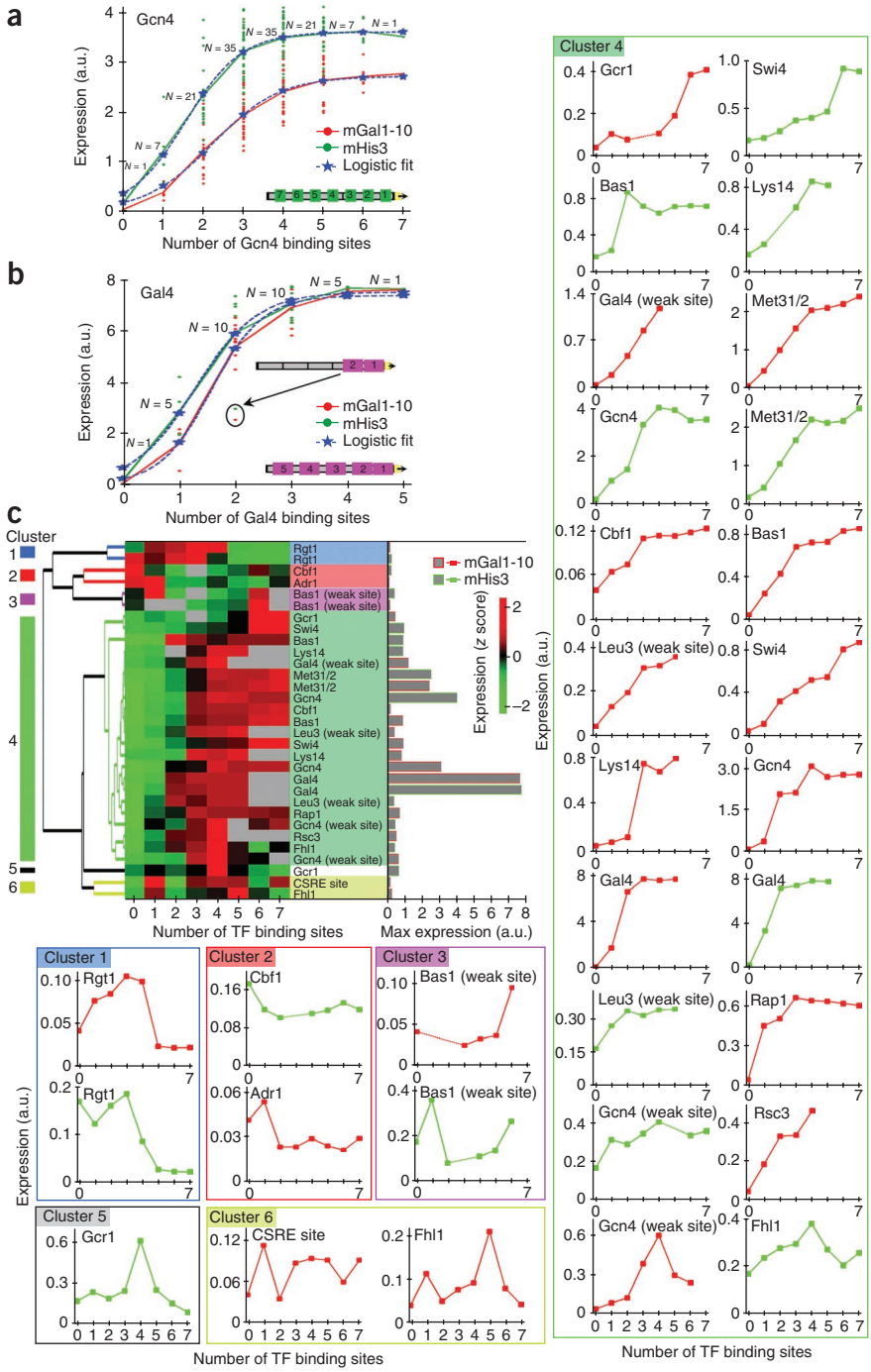

Figure 5 Effect of binding-site number on expression. (**a**) Expression is, on average, a monotonic function of Gcn4 sites that nearly saturates at three or four sites. Within two different promoter backgrounds, we separately inserted Gcn4 sites in all 128 (2^7) possible combinations of sites at seven predefined locations within the promoter. Shown are the individual promoter expressions for each background and mean expressions of all promoters that have *k* Gcn4 sites for *k* = 0, 1, 2..., 7. Also shown is a fit of a logistic function for each background. (**b**) As in **a**, but for all 32 (2^5) possible combinations of Gal4 sites at five predefined promoter locations. The outlier promoter in terms of expression in which the two Gal4 sites closest to the core promoter were both added is indicated. These two sites were added at a distance of 1 bp, as opposed to the 5-bp distance used between all other adjacent sites, thus suggesting steric hindrance between Gal4 sites at this distance. (**c**) For many transcription factors, expression is generally a monotonically increasing function of the number of sites. Hierarchical clustering and heatmap of the expression profile of 31 sets of promoters. In each set, the same transcription-factor site was inserted in *k* copies within the same promoter background for *k* = 0, 1, 2..., 7. Within the heatmap, expression profiles of each transcription-factor site were normalized to a mean of 0 and an s.d. of 1. The 31 sets correspond to 18 different transcription-factor sites (15 different transcription factors, as 3 transcription factors have two site variants differing in their affinity), with each site inserted in two different promoter backgrounds. Absolute expression of the strongest promoter for each transcription-factor site (right, gray bars), showing that the expression at saturation differed greatly among the different sites. We defined six clusters from the hierarchical clustering based on the correlations between the expression profiles of the various transcription factors, and the expression profiles for the individual transcription-factor sites in each cluster are shown within the colored boxes (right and bottom).

combination of transcription factor and promoter background, we found an overall trend of decay in the effect of transcription-factor sites as their distance from the gene start increased, even within the ~100-bp region that could be examined using our approach. However, this trend is relatively weak and does not explain much of the effect of site location on expression.

The effect of nucleosome-disfavoring sequences

Previous studies showed that placing nucleosome-disfavoring sequences, specifically poly(dA:dT) tracts, next to transcription-factor sites significantly affects expression in a manner that depends on the length, composition and location of the tract and is mostly positive regardless of the identity of the transcription factor[8,34]. However, because these findings were derived from dozens of variants of the same promoter background, we sought to test whether they generalize more broadly using the larger number of promoters that can be examined using our method. Notably, using 777 promoters in which we separately inserted consensus sites for 14 transcription factors in two different promoter backgrounds while varying either the site location or the location, length and/or orientation of the poly(dA:dT) tract, we found effects that were consistent with, and thus considerably generalize, previous findings[8,34] (**Fig. 4a–c** and **Supplementary Fig. 16**).

We also explored a previously unknown aspect of poly(dA:dT) tracts by comparing the magnitude of their effect on expression to those of Reb1 and Abf1 sites, as the high nucleosome depletion of these sites *in vivo* was suggested to result from the self action of the transcription factors[15]. Notably, although adding Reb1 and Abf1 sites resulted in a significant increase in expression ($P < 10^{-4}$, $P < 0.02$

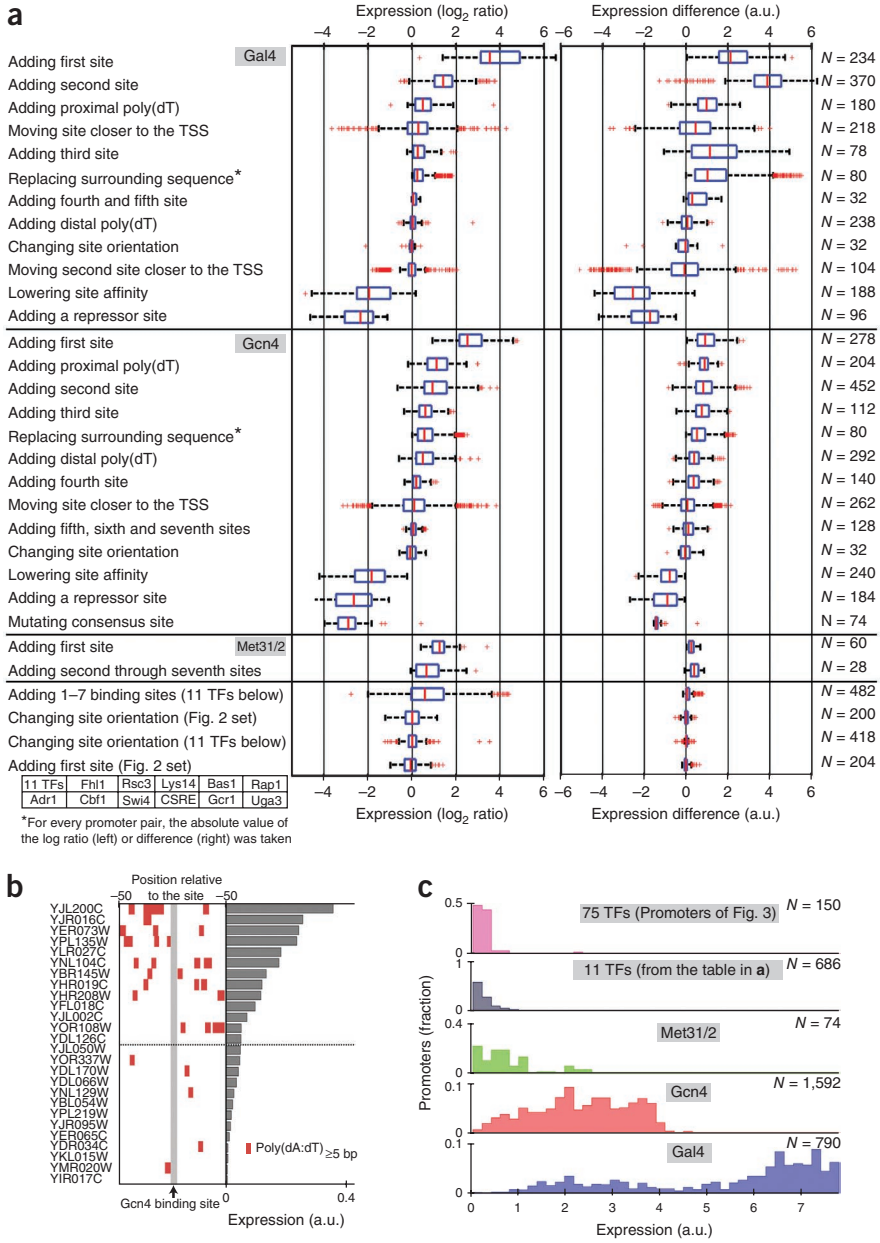

Figure 6 Comparing the effects of different types of sequence changes. (**a**) Effects on expression of different types of sequence changes, either as the change in the log ratio of expression (left) or absolute expression (right). In every row, the boxplot summarizes the effect of a particular type of sequence change (indicated by the text on the left), and each point in the boxplots compares the expression of a promoter in which the change was done to the expression of the same promoter without the change. The first block of changes (12 types) shows changes to Gal4 sites or promoter-containing Gal4 sites, the second block is similar but for Gcn4 (13 types), the third block is similar but for Met31/2 (2 types), and the fourth block (4 types) pulls together the changes to 11 different transcription-factor sites. The number of promoters used in each boxplot is indicated on the right. In each block, the rows are sorted by their effect on the ratio of expression (shown on the left). (**b**) Native yeast promoters with poly(dA:dT) tracts near Gcn4 consensus sites are more highly expressed. Expression (gray bars) of 26 native yeast promoters that contain a consensus Gcn4 site, along with the distribution of poly(dA:dT) tracts at least 5 bp in length in the 100 bp surrounding the Gcn4 site (left, heatmap). Each promoter was measured by the fluorescence of a strain in which it was fused to a YFP reporter, as previously described[19]. The more highly expressed promoters are enriched for poly(dA:dT). (**c**) The expression of promoters with Gal4 or Gcn4 sites is much higher than that of all promoters with sites for other transcription factors. Distributions of expression for five different promoter sets, showing promoters with single sites for 75 different transcription factors (first row); promoters with various manipulations to sites for 11 different transcription factors, including promoters with up to seven sites for each of these transcription factors (second row); and all of the promoters that contain only Met31/2 sites (third row), Gcn4 sites (fourth row) or Gal4 sites (fifth row). The last three rows include all of the manipulations to promoters with sites for these transcription factors.

respectively), the effect was comparable to that of adding a 10-bp poly(dA:dT) tract and was significantly less than that of adding a 15-bp tract ($P < 10^{-6}$; **Fig. 4d**). These results suggest that promoter-driven expression can be increased to similar amounts by depleting nucleosomes with either the *cis*-regulatory mechanism of poly(dA:dT) tracts or the *trans*-regulatory mechanism of sites for general transcription factors, such as Reb1 and Abf1.

The effect of binding-site number

We next used our ability to design promoters with multiple combinations of transcription-factor sites to systematically test the dependence of reporter gene expression on the number of sites. We selected two promoter contexts, and for each context, we separately inserted the consensus sites for Gcn4 and Gal4 in all 128 (2^7) and 32 (2^5) possible combinations of sites at seven and five predefined locations, respectively. Notably, we found a clear relationship between the number of sites and the average expression from the promoters with that number of sites for both transcription factors in both contexts, which accurately fits a logistic function ($R^2 = 0.99$; **Fig. 5a**,**b**). In all cases, expression increased with the addition of each site up to and including three to four sites but then mostly saturated.

Despite this close fit of the average expression of a given number of sites to a logistic function, individual promoters with specific combinations of site locations deviated from the expression predicted for them by this logistic model. Part of this deviation probably stems from the different effects that sites have at different promoter locations, whereas another part probably results from nonadditive interactions between pairs of sites, predominantly from interactions between adjacent sites (**Supplementary Figs. 17** and **18**). Notably, our results suggest that two Gal4 molecules sterically occlude each other in binding to two sites whose ends are 1 bp apart and that Gcn4 may have similar albeit weaker behavior when its site ends are 5 bp apart (**Supplementary Fig. 19**).

We extended the above set to 13 additional transcription factors at lower resolution, and for each transcription factor we generated promoters with zero, one or up to five (1 transcription factor) or seven

(12 transcription factors) binding sites in increments of one site in two different background sequence contexts. At this lower resolution, the results are more sensitive to location-specific site contributions, as there is only one promoter for each transcription factor in every combination of context and site number. Nevertheless, we observed clear trends: for most transcription factors, expression largely increases as the number of binding sites increases, mostly saturating at around three or four sites (**Fig. 5c** and **Supplementary Fig. 20**). One notable exception is Rgt1, for which expression is a nonmonotonic function of site number, typically increasing with the first three sites but then markedly decreasing at four or more sites (**Fig. 5c** and **Supplementary Fig. 21a**). This suggests that Rgt1 is a potent repressor only with more than four sites, which is consistent with a previous study of one native Rgt1 target[35]. For the Matα2p-Mcm1p repressor, we also found stronger repression with more sites, although in this case, repression was already evident with only one site (**Supplementary Fig. 21b**).

Thus, we found a clear relationship between gene expression and the number of activator sites that accurately fits a logistic function in which expression increases monotonically with more sites and then mostly saturates at around three or four sites. Notably, the expression at the saturation point differed greatly among transcription factors, and, with one exception (Met31/2), all of the promoters for the 11 transcription factors we tested, including those with seven sites, had much lower expression than that of a promoter with a single Gal4 site or one or two Gcn4 sites (**Fig. 5c**). This suggests that in the growth condition and promoter backgrounds we tested, the transcription-factor binding-site identity is more crucial than site number for achieving high expression.

Comparing different sequence changes

Finally, to obtain a high-level view of our library, we partitioned the 6,500 promoters into sets, where each set represented changes to the same type of regulatory element. Within Gal4- and Gcn4-regulated promoters, we found coherent behavior that may be expected from previous research; in most cases, adding sites for these transcription factors or adding poly(dA:dT) tracts increased expression, and lowering site affinity or adding repressor sites decreased expression (**Fig. 6a**). By contrast, increasing the distance of Gal4 or Gcn4 sites from the start of the gene had stimulatory effects in some cases and inhibitory effects in others. The different types of sequence changes also showed a fairly robust ranking in the magnitude of their effect, with the largest effect coming from addition of the first one/two Gal4 or Gcn4 sites or of a proximal poly(dA:dT) tract to a promoter that contained at least one Gal4 or Gcn4 site (**Fig. 6a**). To test the applicability of the observation that poly(dA:dT) strongly increase expression in endogenous promoters, we generated fluorescent reporter strains for 26 yeast promoters with a consensus Gcn4 site and indeed found a significant enrichment of poly(dA:dT) tracts in the more highly expressed promoters ($P < 0.003$; **Fig. 6b**).

Notably, the expression of all 836 promoters in which we manipulated sites for 75 transcription factors other than Gal4 and Gcn4 was markedly lower than the vast majority of the 602 promoters that contained just a single Gal4 or Gcn4 site (**Fig. 6a**,**c**). These 836 promoters represent a variety of changes to the location and orientation of transcription-factor sites and, for 11 transcription factors, include promoters with one, two or up to seven sites. Although Gal4 and Gcn4 are activated in the growth condition we used here (galactose medium starved for amino acids), the magnitude of the expression difference is notable. The reason for this finding is unclear. Possible explanations include higher amounts of active Gal4 and Gcn4 molecules, stronger activation domains or the tested promoter contexts being less suitable for the other transcription factors. Regardless of the reason, our results suggest that, at least in our tested condition and contexts, transcription-factor identity is the most crucial factor in achieving high expression.

DISCUSSION

We present in this study a high-throughput method for measuring the expression of thousands of fully designed promoters in a single experiment and with accuracy comparable to that obtained when the promoters are constructed and measured individually. We applied our method to study how expression depends on various parameters such as the identity, number, affinity and location of transcription-factor binding sites, enabling a large-scale systematic testing of the effects of these parameters. For several types of sequence manipulations, our data reinforce previous results or support hypotheses that have arisen from smaller-scale studies (**Supplementary Note 1**). For other types of manipulations, the effects were more unexpected, and their mechanistic bases are unclear, raising questions for further research. For example, we found that changing a transcription-factor site location by even a few base pairs typically exerts large effects on gene expression. We were also surprised by the higher expression of most of the 602 promoters with even a single Gal4 or Gcn4 site compared to the expression of all ~700 promoters that contained sites for 11 other transcription factors. Notably, these ~700 promoters include nucleosome-disfavoring sequences and up to seven sites for each of these transcription factors. Moreover, even when the qualitative effects matched our expectations, the next challenge will be to mechanistically explain the quantitative magnitude of the effects. Despite the above insights, our method has several limitations, the most notable of which is the limited (~100-bp) length of the promoter region that we could vary (**Supplementary Note 2**).

For decades, researchers have sought a 'regulatory code' that translates DNA sequence into expression. The fact that several types of sequence changes that we performed here have predictable effects on expression that hold across many contexts and transcription factors suggests that such a general code may indeed exist, but because of the many unexplained effects that we found, it is also clear that we are far from deciphering it. The ability to carefully design large-scale promoter libraries should prove useful in improving our understanding of the regulation of transcription in eukaryotes, eventually leading to quantitative predictive models of transcriptional regulation. It will also be exciting to apply similar strategies to study the effects that other regulatory layers, such as post-transcriptional regulation, codon usage and translation initiation regulation, have on gene expression and other biological phenotypes.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession code. GEO: GSE37851 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37851).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Widom for assistance and inspiration throughout this project. This work was supported by grants from the European Research Council and the US National Institutes of Health to E. Segal. E. Segal is the incumbent of the Soretta and Henry Shapiro career development chair. We thank S. Lubliner for help with computational analyses. We thank C. Boone (University of Toronto) for kindly giving us the Y8205 strain.

AUTHOR CONTRIBUTIONS

E. Sharon and E. Segal conceived the project. E. Sharon., Y.K., A.W. and E. Segal planned the experiments. E. Sharon and Y.K. performed the experiments. E. Sharon and E. Segal analyzed the results. T.R.-S., M.L. and Z.Y. contributed to the design of the promoters. A.S., D.Z. and L.K. contributed to experimental work. Z.Y. also provided technical guidance.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nbt.2205.

1. Chiang, D.Y., Nix, D.A., Shultzaberger, R.K., Gasch, A.P. & Eisen, M.B. Flexible promoter architecture requirements for coactivator recruitment. *BMC Mol. Biol.* **7**, 16 (2006).
2. Ligr, M., Siddharthan, R., Cross, F.R. & Siggia, E.D. Gene expression from random libraries of yeast promoters. *Genetics* **172**, 2113–2122 (2006).
3. Kinkhabwala, A. & Guet, C.C. Uncovering *cis* regulatory codes using synthetic promoter shuffling. *PLoS ONE* **3**, e2030 (2008).
4. Gertz, J., Siggia, E.D. & Cohen, B.A. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* **457**, 215–218 (2009).
5. Cox, R.S. III., Surette, M.G. & Elowitz, M.B. Programming gene expression with combinatorial promoters. *Mol. Syst. Biol.* **3**, 145 (2007).
6. Kinney, J.B., Murugan, A., Callan, C.G. Jr. & Cox, E.C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. USA* **107**, 9158–9163 (2010).
7. Giniger, E. & Ptashne, M. Cooperative DNA binding of the yeast transcriptional activator GAL4. *Proc. Natl. Acad. Sci. USA* **85**, 382–386 (1988).
8. Iyer, V. & Struhl, K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* **14**, 2570–2579 (1995).
9. Lam, F.H., Steger, D.J. & O'Shea, E.K. Chromatin decouples promoter threshold from dynamic range. *Nature* **453**, 246–250 (2008).
10. Murphy, K.F., Balazsi, G. & Collins, J.J. Combinatorial promoter design for engineering noisy gene expression. *Proc. Natl. Acad. Sci. USA* **104**, 12726–12731 (2007).
11. Patwardhan, R.P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
12. Patwardhan, R.P. *et al.* Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nat. Biotechnol.* **30**, 265–270 (2012).
13. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
14. LeProust, E.M. *et al.* Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38**, 2522–2540 (2010).
15. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2009).
16. Baliga, N.S. Promoter analysis by saturation mutagenesis. *Biol. Proced. Online* **3**, 64–69 (2001).
17. Anderson, J.D. & Widom, J. Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. *Mol. Cell. Biol.* **21**, 3830–3839 (2001).
18. Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.* **19**, 65–71 (2009).
19. Zeevi, D. *et al.* Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters. *Genome Res.* **21**, 2114–2128 (2011).
20. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
21. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
22. Huh, W.K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
23. Zhao, Y. *et al.* Fine-structure analysis of ribosomal protein gene transcription. *Mol. Cell. Biol.* **26**, 4853–4862 (2006).
24. Blaiseau, P.L., Lesuisse, E. & Camadro, J.M. Aft2p, a novel iron-regulated transcription activator that modulates, with Aft1p, intracellular iron use and resistance to oxidative stress in yeast. *J. Biol. Chem.* **276**, 34221–34226 (2001).
25. Lamb, T.M. & Mitchell, A.P. The transcription factor Rim101p governs ion tolerance and cell differentiation by direct repression of the regulatory genes NRG1 and SMP1 in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **23**, 677–686 (2003).
26. Hanlon, S.E., Rizzo, J.M., Tatomer, D.C., Lieb, J.D. & Buck, M.J. The stress response factors Yap6, Cin5, Phd1, and Skn7 direct targeting of the conserved co-repressor Tup1-Ssn6 in *S. cerevisiae*. *PLoS ONE* **6**, e19060 (2011).
27. Canizares, J.V., Pallotti, C., Sainz-Pardo, I., Iranzo, M. & Mormeneo, S. The SRD2 gene is involved in *Saccharomyces cerevisiae* morphogenesis. *Arch. Microbiol.* **177**, 352–357 (2002).
28. Akache, B., Wu, K. & Turcotte, B. Phenotypic analysis of genes encoding yeast zinc cluster proteins. *Nucleic Acids Res.* **29**, 2181–2190 (2001).
29. Woudt, L.P., Smit, A.B., Mager, W.H. & Planta, R.J. Conserved sequence elements upstream of the gene encoding yeast ribosomal protein L25 are involved in transcription activation. *EMBO J.* **5**, 1037–1040 (1986).
30. Lieb, J.D., Liu, X., Botstein, D. & Brown, P.O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat. Genet.* **28**, 327–334 (2001).
31. Nutiu, R. *et al.* Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.* **29**, 659–664 (2011).
32. Maerkl, S.J. & Quake, S.R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).
33. Bulyk, M.L., Gentalen, E., Lockhart, D.J. & Church, G.M. Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nat. Biotechnol.* **17**, 573–577 (1999).
34. Raveh-Sadka, T. *et al.* Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.* (in the press).
35. Kim, J.H., Polish, J. & Johnston, M. Specificity and regulation of DNA binding by the yeast glucose transporter gene repressor Rgt1. *Mol. Cell. Biol.* **23**, 5208–5216 (2003).

ONLINE METHODS

Library design. We designed a total of 6,500 promoters, divided into several sets, each aimed at examining a specific aspect of transcriptional regulation. We designed the library using a 'LEGO'-like methodology: each promoter is composed from a context DNA sequence and several regulatory elements, such as transcription-factor binding sites and nucleosome-disfavoring sequences. With a few exceptions (described below), the elements are integrated into the promoter by replacing the background DNA at the relevant position. Because we were interested in the position of the regulatory elements relative to the YFP gene, our coordinates are specified from the 3′ end of the oligo to the 5′ end and are one based (3′ most bp is coordinate 1; coordinates increase in upstream direction). Correspondingly, the start of each element is its 3′ end position. In several cases, regulatory elements were integrated by insertion. In these cases, the insertion position is described in the library design file by a noninteger coordinate. The integer portion of the number specifies the number of background nucleotides downstream of the insertion, and the fraction indicates the order of the inserted element from 3′ to 5′ (in case several elements were inserted into a similar position). When regulatory elements were integrated into the background sequence by insertion rather than by replacement, the sequence is truncated in its 5′ end to maintain uniform oligo lengths. The library description file contains a unique identifier for every promoter, a description text that specifies the background DNA (referred to as 'context' DNA), and the sequence and location of regulatory elements, barcode and PCR primers (for technical reasons, the –1 coordinate indicates that the element does not exist in the promoter) and the library oligo sequence (**Supplementary Tables 1**–**3**).

Background sequences. The library is constructed with two primary background DNA sequences, termed mGal1-10 (in the description files "GAL1-10_NULL") and mHis3 (in the description files "HIS3_NULL"). The mGal1-10 background is a sequence from the native yeast Gal1-10 promoter in which the known regulatory elements were mutated. The mHis3 background is taken from the yeast His3 promoter with known regulatory elements deleted. Coding ("CDS"), intergenic and promoter backgrounds were selected randomly from the corresponding region in the yeast genome, as annotated in the UCSC genome browser[36]. The "GC40Random*" backgrounds were generated randomly with a mean 40% dG/dC. "<gene name>_NATIVE" are unmodified DNA sequences from the specified gene promoter[36]. The "GAL1_10_GINIGER*" sequences are variations of the yeast Gal1-10 promoter.

Regulatory elements. For each transcription factor, we selected one or more binding sites based on the literature cited in the regulatory elements description file. Our regulatory elements also contain nucleosome-disfavoring sequences (indicated by type "Boundary"). The main nucleosome-disfavoring sequence that was used in the library is a stretch of poly(dT) of length 15 bp.

Design of promoter barcodes. The 5′ edge of each designed promoter contains a unique 10-bp sequence, which we used as a barcode for the promoter that allowed us to uniquely identify it using a short sequencing read. We designed the barcode sequences such that every barcode differs from the any other barcode in at least 3 bp, allowing us to correctly identify every barcode even if it contains a single–base-pair mutation. We excluded low-complexity sequences from being barcodes, as they may affect the PCR reactions. To minimize the effect of the barcode on promoter expression, we also excluded sequences that resemble any known yeast transcription-factor binding-site sequence by ensuring that they do not match known published consensus sequences[20,37].

Design of expression bin barcodes. To identify the expression bin of each promoter from the multiplexed sequencing lane, we added a unique 5-bp tail to the 5′ end of the primer that is used to amplify the synthetic promoter region of every promoter expression bin. We chose the bin barcodes such that every barcode differed from the other in the last 2 bp, such that a single–base-pair mutation cannot cause an error in the bin mapping. We excluded low-complexity sequences from being barcodes, as they may affect the PCR reactions.

Synthetic promoter library preparation. To prepare a large-scale library of 6,500 different types of promoters, each fused to a fluorescent reporter and with each cell containing one of the 6,500 synthetic promoters, we first used Agilent Oligo library synthesis technology to produce a pool of 6,500 different single-stranded 150-mers. The library is synthesized on programmable microarrays by Agilent[14,38] and then provided to us as an oligo pool in a single tube (10 pmol). Each oligo contains common priming sites and restriction sites at both ends, as well as a 10-bp unique barcode, leaving 103 bp for the variable promoter region in which we performed the various library manipulations. The pool was dissolved in 200 µl Tris-EDTA. We divided 12 ng (2.6%) of the single-stranded library DNA into 96 wells for PCR amplification in a final volume of 50 µl. Each reaction contained 24 µl of water containing 0.125 ng DNA, 10 µl of 5× Herculase II reaction buffer, 10 µl 2.5 mM dNTPs mix, 2.5 µl 20 µM 5′ primer, 2.5 µl 20 µM 3′ primer and 1 µl Herculase II Fusion DNA Polymerase (Agilent[39]). The parameters for PCR were 95 °C for 1 min, 12 cycles of 95 °C for 20 s and 68 °C for 1 min and then one cycle of 68 °C for 4 min. The primers used to amplify the library were 5′-GGGGACCAGGTGCCGTAA-3′ (forward primer) and 5′-TGATCGCCCTAGGATCGC-3′ (reverse primer). The PCR products from all 96 wells were joined and concentrated using Amicon Ultra, 0.5 ml 30K centrifugal filters for DNA purification and concentration. The concentrated DNA was then purified using a PCR MiniElute Purification Kit (QIAGEN) according to the manufacturer's instructions.

Construction of the library plasmid. Plasmid pKT103 (ref. 40), which contains yEVenuse, was used as a vector backbone to create the recipient plasmid (pPAL1_His3core) for the library. The KAN cassette was replaced with a NAT cassette. We amplified TEF2 promoter–mCherry from the pAG60-TEF2-Cherry plasmid[41] and inserted it as an EcoO109I/AatII fragment into pKT103. We amplified URA3 from plasmid pRS316 and inserted it to the recipient plasmid. To allow for chromosomal segregation of the library plasmids, we introduced an ARS-CEN sequence. The ARS-CEN sequence was amplified from the pRS316 plasmid and inserted as an EcoRI/EcoRV fragment. We amplified the first 100 bp upstream of the genomic His3 ATG, which served as a core promoter for the library, and inserted it upstream of the yEVenus ATG. The His3 core promoter was flanked by a sequence with two restriction sites that were used for the ligation of the 150 bp pooled library.

Ligation and transformation. Purified library DNA (150 ng) was cut with the restriction enzymes AvrII and SexAI (Fermentas) for 2 h at 37 °C in a reaction mixture containing 3 µl FD buffer (supplied by Fermentas), 0.8 µl of each enzyme and 25.4 µl DNA. Digested DNA was separated from the smaller fragments by electrophoresis on a 2.5% agarose gel stained with ethidium bromide. Only fragments of the correct size (128 bp with a 9-bp overhang) were cut from the gel and eluted using electroelution Midi GeBAflex tubes. Eluted DNA was precipitated using the standard sodium acetate and isopropanol protocol. To prepare the vector for cloning, the plasmid was cut with the restriction enzymes AvrII and SexAI for 2 h at 37 °C in a reaction mixture containing 6 µl FD buffer, 3 µl of each enzyme, 3 µl alkaline phosphatase (Fermentas) and 3 µg of the plasmid in a total volume of 60 µl. Digested DNA was purified using a PCR purification kit (QIAGEN). The digested plasmid and library DNA were ligated for 30 min at room temperature in a 10 µl reaction that contained 150 ng plasmid, 2.8 ng library DNA (molar ratio of 1:1 plasmid to library DNA), 1 µl CloneDirect 10× Ligation Buffer and 1 µl CloneSmart DNA Ligase (Lucigen Corporation). The library DNA was thus ligated to the plasmid directly upstream of a short core promoter (100 bp TATA-containing core promoter taken from the native His3 yeast promoter) that is followed by a yEVenus gene. Ligated DNA was transformed into seven tubes, each containing 25 µl of *E. cloni* 10G electrocompetent cells (Lucigen), which were then plated on 28 15-cm plates containing lysogeny broth (LB) and ampicillin. Sixteen hours after transformation, the plates containing 50,000 colonies each were scraped into LB medium, and the plasmid DNA was purified using a plasmid maxi kit (QIAGEN). To minimize the number of plasmids with multiple inserts, we cut the plasmid only with the SexAI restriction enzyme, ran it on a gel stained with crystal violet, purified the digested plasmid from the gel with a QIAquick gel purification kit (QIAGEN) and then re-ligated the plasmid using CloneSmart DNA Ligase. Transformation was performed as described above. Cells from six transformations were plated on 30 15-cm plates. Colonies were scraped (84,000 per plate), and plasmids were purified with a plasmid maxi kit (QIAGEN).

Yeast transformation. To achieve a high efficiency of transformation, we used the electroporation protocol[42] to transform the library plasmid into the Y8205 strain (kind gift from C. Boone). For a single transformation, we grew the culture to an optical density (OD_{600}) of 1.4, 10^8 cells were washed twice with cold double deionized water and once with cold 1 M sorbitol. Cells were suspended in 40 µl cold 1 M sorbitol. Library plasmid (0.5 µg) was mixed with the cells. The mixture was kept on ice for 5 min and transferred to an electroporation cuvette (0.2 mm), and a 1.5-kV, 25-µF, 200-Ω pulse was applied (Gene Pulser, Bio-Rad, Richmond, California, USA). The cells were diluted immediately with 0.95 ml of SCD-URA (synthetic complete media with 2% glucose and without uracil) medium and were then diluted to 1:20, to reach final OD_{600} of 1, and transformed cells (120,000 transformants) were grown at 30 °C for 72 h until the culture reached the stationary phase. To preserve the diversity of the library, we performed this transformation 23 times.

Sorting by flow cytometry. To adapt cells to the medium of the sorting, stationary yeast library cells were diluted in SC-Gal-URA (synthetic complete media with 2% galactose and without uracil) medium without amino acids, except for histidine and leucine, and were grown to the stationary phase. Next, this culture was diluted again in similar medium (to OD_{600} 0.03–0.05) and grown to the mid-exponential phase (OD_{600} 0.5–1.5) for sorting. Sorting was performed with the FACSAria cell sorter (Becton-Dickinson) at the low sample flow rate and a sorting speed of no more than 20,000 cells per s. To reduce the effect of extrinsic noise on promoter activity, we sorted only cells that were gated to have relatively homogeneous size and mCherry fluorescence (corresponding to approximately one or two plasmid copies). The library was sorted using two strategies. In one strategy (replicate 1), the cells were sorted three times recursively into four bins, producing a total of 64 bins (we first sorted all the cells into four bins, applied the same sorting procedure to the cells in each bin and then applied this procedure again to the cells of each bin sorted at this second level). In another strategy (replicate 2), we sorted the cells directly into 16 bins. We chose this number of bins as a compromise between sorting time and having enough bins to resolve the expression to a good resolution (shown by the excellent agreement of $R^2 = 0.99$ that we obtained between our expression measurements and those of the isolated strains). In both strategies, cells were sorted according to the ratio of YFP and mCherry (located in the plasmid with a promoter that is constant across the library, the yeast TEF2 promoter), thereby normalizing for extrinsic noise effects. Because the mCherry reporter is driven by the same promoter across the entire library, sorting by the ratio of YFP to mCherry corrects for extrinsic noise across the cell population. The distribution of mCherry expression intensities peaked at several discrete values, corresponding to the different number of plasmids integrated in each cell. The first peak was substantially higher than all the others, suggesting that the vast majority of cells had a single plasmid (likely because of our use of low–copy-number plasmids and a low concentration of plasmid DNA during the transformation process), and we thus gated the cells by the mCherry expression of the first peak to enrich for this population. In each sorting strategy, the expression bins contained equal fractions of the library cells, and we collected a total of 8,000,000 cells. We note that because both YFP and mCherry are stable and long lived[19], their fluorescence integrates their expression over several cell cycles, and thus, promoters that are expressed only in a specific cell-cycle phase will have lower expression than those that are constitutively expressed throughout the cell cycle. For these reasons, we made minimal use of cell-cycle–dependent regulators.

Isolating control strains. To obtain isolated strains from the library, a small aliquot of cells from each bin was plated on SCD-URA medium. Six colonies from each bin were picked and further grown in a selective medium and were used for sequencing and measurements. The PCR product containing the promoter from 96 individual colonies (3 from each bin) was sent to Sanger sequencing. Promoter activities from all the sequenced stains were measured using both flow cytometry (with an LSRII cell analyzer and a similar protocol to that used for the library measurements) and a plate reader (Tecan Infinite F500) as previously described[34].

Preparing samples for sequencing. Sorted cells were grown in 5 ml SCD-URA medium to stationary phase. One million cells from each bin were taken for colony PCR using specific primers corresponding to the promoter region of the plasmid. The 3′ primer was common to all bins (5′-NNNNNTTATGTGATAATGCCTAGGATCGC-3′, where the Ns represent random nucleotides). The 5′ primer had a common sequence and a unique upstream 5-bp barcode sequence (underlined) that was specific to each bin (5′-XXXXXGGGGACCAGGTGCCGTAAG-3′, where the Xs represent the bin's unique sequence). In replicate 2, DNA from each bin was subjected to PCR with two different 5′ primers. PCR products were purified using ZR-96 DNA Clean and Concentrator-5 (Zymo Research). An equal amount of DNA from each PCR product (2.35 ng) was joined to one tube and concentrated using Amicon Ultra 0.5 ml 30K centrifugal filters, and the sample (100 ng) was then sent to sequencing by the SOLiD system.

Sequencing. We sequenced the sample using Applied Biosystem's SOLiD high-throughput sequencing platform (AB SOLiD) and mapped it to a reference sequence set that contained all combinations of strain and expression bin barcode sequences using SHRiMP[3]. We obtained ~13,000,000 (replicate 1) and ~14,000,000 (replicate 2) uniquely mapped reads that each contained a strain and expression bin barcode pair.

Deriving mean promoter expressions. We first obtained the mean and s.d. of the expression of each sorted flow cytometer bin, which we estimated from the expressions of the cells sorted into each bin. Next, for each promoter, we estimated the corresponding fraction of cells in each bin based on the number of sequence reads from that bin that mapped to that promoter (the reads of each bin were first normalized to match the fraction of the bin in the entire population). We then defined the mean expression of each promoter as a weighted average of the mean expression of all bins, where the weight of each bin is the fraction of the promoter in that bin. Similarly, we computed the s.d. of each promoter using the s.d. of each bin and the distribution of the promoter across the bins. When our experiment consisted of more than one round of sorting, we used the results of the last round, as described above.

Deriving expression predictions from *in vitro* binding affinities. To compare our expression measurements of the effects of mutations of Gcn4 binding sites (**Supplementary Fig. 6b**) to the *in vitro* binding affinities of Gcn4 sites, we used a simple model to derive expression predictions from the *in vitro* affinities. To this end, we assumed that expression is linearly related to the promoter occupancy of Gcn4 and computed Gcn4 occupancy as a function of the binding-site dissociation constant (K_d) using a simple Michaelis-Menten–based model, adding a term that accounts for the extra observed activity of Gcn4 sites in the reverse orientation. Thus, given the measured dissociation constant of a Gcn4 site, the predicted expression is given by:

$$\text{expression} = a*(1+b*\delta)*\frac{[\text{Gcn4}]}{[\text{Gcn4}]+K_d}$$

where a, b and [Gcn4] are the free parameters of the model (scaling factor, orientation factor and Gcn4 concentration, respectively) and δ is an indicator function for whether the Gcn4 site is in reverse orientation relative to His3 native Gcn4 site. We fit the free parameters so as to best fit our measured expressions.

Visualizing comparisons of expression across promoter sets. Rather than using traditional boxplots, we used the MATLAB function "notBoxPlot", written by R. Campbell, to visualize various comparisons of expression values across different promoter sets.

Statistical analyses. To ensure that our various results are robust to outliers, we adapted robust statistical practices and used the median instead of the mean as an estimator of the central tendency and the median absolute deviation scaled to the normal distribution dispersion as an estimator of the s.d.[43]. To assess the effects of various nucleosome-disfavoring elements on expression, we used a two-sided two-sample t test. To show that the effect of the surrounding sequence on the activity of regulatory elements is significantly larger than the barcode effect on strains with an identical promoter, we used a two-sample F-test for equal variance between each two promoter groups (using the MATLAB function "vartest2").

doi:10.1038/nbt.2205

36. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
37. Zhu, C. *et al.* High-resolution DNA binding specificity analysis of yeast transcription factors. *Genome Res.* **19**, 556–566 (2009).
38. Cleary, M.A. *et al.* Production of complex nucleic acid libraries using highly parallel *in situ* oligonucleotide synthesis. *Nat. Methods* **1**, 241–248 (2004).
39. Fazekas, A., Steeves, R. & Newmaster, S. Improving sequencing quality from PCR products containing long mononucleotide repeats. *Biotechniques* **48**, 277–285 (2010).
40. Sheff, M.A. & Thorn, K.S. Optimized cassettes for fluorescent protein tagging in *Saccharomyces cerevisiae*. *Yeast* **21**, 661–670 (2004).
41. Breslow, D.K. *et al.* A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat. Methods* **5**, 711–718 (2008).
42. Otsuka, C. *et al.* Use of yeast transformation by oligonucleotides to study DNA lesion bypass *in vivo*. *Mutat. Res.* **502**, 53–60 (2002).
43. Hoaglin, D.C., Mosteller, F. & Tukey, J.W. *Understanding Robust and Exploratory Data Anlysis* (Wiley, 1983).