Supplementary information for

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

Supplemental note 1. Overview of main biological findings.

Our measurements represent the first large-scale systematic testing of the effects of several different factors on expression. For several types of sequence manipulations, our data reinforce previous results or support hypotheses that have arisen from smaller scale studies. These include (1) the general stimulatory effect on expression that we found when adding nucleosome disfavoring poly(dA:dT) tracts; (2) the increase and decrease in expression that results in nearly all cases in which activator and repressor sites are added, respectively; (3) the decrease in the contribution of adding an activator TF site as the number of previously existing sites in the promoter increases; and (4) a clear relationship that we found between expression and the number of TF binding sites, which can be described with high accuracy by a logistic function in which expression mostly saturates at 3-4 sites.

We note, however, that in all of the above cases, our data considerably generalizes previous results (typically based on handful of promoters) to many more promoters, sequence backgrounds, and transcription factors. More importantly, the thousands of promoters in which we performed the above manipulations provide a highly valuable resource that should allow us to go beyond these qualitatively expected behaviors and to the next unresolved challenge of explaining the quantitative magnitude of the effects that are accurately measured by our method.

In other cases, where our results also provide a clear measure of the effect of the tested sequence manipulation, the effect on expression is more surprising and its mechanistic basis is unclear, raising interesting open questions for further research. These include (1) the finding that small changes of even a few basepairs in the location of TF binding sites typically have large effects on expression; (2) the finding of a decay of the effect of transcription factor sites as their distance to the gene start increases, whereby repressor sites that are moved further away from the gene start result in higher expression and activator sites that are moved further away results in lower expression; (3) the novel finding of a ~10bp periodic relationship between expression and the distance of Gcn4 sites from the gene start; and (4) the dramatically higher expression that we found for Gal4 and Gcn4 regulated promoters, even those that contain a single site for these factors, compared to the expression of all ~700 promoters that contained sites for 11 other TFs. Notably, these ~700 promoters included additions of nucleosome disfavoring sequences and of up to seven sites for each of these TFs in two different contexts, suggesting that at least in our tested condition and context, no other manipulations to sites for these TFs can achieve the levels of expression of promoters with Gal4 and Gcn4 sites.

In all of these latter cases, the mechanistic basis is unclear. However, we believe that since each of these patterns are based on at least hundreds promoters, the phenomenon that we describe is real and thus in these cases our results point to concrete and interesting open questions for further research as well as to testable hypotheses. Regarding (1) above, while the effect of TF sites generally decayed with their distance from the core promoter, this decay only explains a small fraction of the effect of site location, leaving open the mechanistic basis for the observed effect. Since we did not anticipate such large effects of small changes in site location, we changed site locations in 3-4bp or 7bp increments, and thus, performing similar analyses at 1bp increments is likely to be a fruitful direction for improving our understanding of the effect of site location.

Regarding (2) above, although perhaps somewhat expected, we are not aware of other large-scale systematic testings in which transcription factor binding sites were moved in small increments over a ~100bp region and an overall trend was seen in the expression output. This trend is particularly noticeable in our data in the case of moving repressors sites, and here too, the mechanism by which expression increases, i.e., the mechanism by which the quanching effect of repressors appears to be decreasing with their distance is unclear.

Regarding (3) above, the ~10bp periodicity that we found between expression and Gcn4 site location was not observed for Gal4 and Leu3, whose sites we also varied at similar 3-4bp increments within the same promoter context, and the pattern is thus unlikely to be due to obscure Idiosyncrasies within the background promoter. Our current intriguing and testable hypothesis is that Gcn4 requires a specific alignment with the transcriptional machinery, which it would reassume at every complete helical turn of the DNA.

Regarding (4) above, although we purposely chose the growth condition (galactose medium starved for amino acids) such that Gal4 and Gcn4 would be activated, the magnitude of the expression differences between their promoters and all others are still surprising and the reason for them is unclear. Here too, there are several testable hypotheses including higher amounts of active Gal4 and Gcn4 molecules, stronger activation domains for these TFs, or that the tested promoter contexts are less suitable for the other tested TFs that we tested. It will be of great interest to decipher all of the above mechanisms in future work.

Supplemental note 2. Limitations of our method.

Despite the insights afforded by our method, our approach has several limitations. Most notably, due to the DNA synthesis technology that we employed, the length of the promoter region that we varied was limited to 103bp. Since the core promoter region was not included in this variable region, a second limitation is the use of the same fixed core promoter across our entire library. If the TFs that we used have differential interactions with core promoters, this may affect the relative expression ranking of our library promoters. Third, since we used most of the real estate of the library to change the sequence along axes such as site location and site number, we could not comprehensively examine other axes such as surrounding sequence, and we thus tested most of our library in only two different promoter backgrounds. It will be interesting to see the robustness of the results when performing the same variation in regulatory elements across many different contexts. Finally, although we would like to interpret the results in terms of binding of transcription factors, by using expression and not binding as the readout we are really only testing the effect of the DNA elements that we manipulate (e.g., transcription factor binding sites).

Supplemental figure legends

Supplemental Figure 1. Our method obtains highly reproducible measurements of expression noise. We isolated 92 individual strains from our pool of transformed yeast cells and sequenced each of them to reveal their identity. Shown is a comparison of cell-to-cell expression variability (noise) measurements (log-scale, standard deviation divided by mean expression) obtained for these strains when each strain was measured in isolation using a flow cytometer (x-axis) or within a single experiment using our method (y-axis).

Supplemental Figure 2. Barcodes have little effect on our expression measurements. Same as Figure 1C, but for 14 additional pairs of promoters that differ only in their barcode sequence, shown is the distribution of sequencing reads across the expression bins.

Supplemental Figure 3. Promoter expression is highly similar whether on plasmids or genomically integrated. Shown is a comparison of the expression that our method obtains for 29 strains (y-axis) against individual expression measurements of strains in which the same promoters were genomically integrated into a fixed location in the yeast genome. Notably, for technical reasons, the genomically integrated strains contain small 10-30bp insertions upstream to the integrated promoter, yet still give highly similar expression values as measured with our pooled method.

Supplemental Figure 4. Identifying functional elements within promoters using systematic scanning mutations. **(A)** For a 103bp taken from the well-studied Gal1/10 promoter from yeast, each bar shows the effect of randomly mutating the three underlying basepairs. The value of each bar is the ratio between the activity of the mutated promoter and that of the original unmodified promoter. Bars are colored by their overlap with putative regulatory elements (green, putative TF binding sites; blue, no known regulatory elements; dark red, poly(dA:dT) tract), with the location and identity of the putative elements marked along the promoter. **(B)** Same as (A), for a 103bp region taken from the native TSA1 yeast promoter. **(C)** Same as (A), for a 103bp region taken from the native RPL3-10 yeast promoter.

Supplemental Figure 5. Assessing the significance of the effect of site orientation on expression. **(A)** Effect of TF site orientation on expression. Shown is a ranking of all 75 tested TFs according to the ratio between the expression of the orientation of their site with higher expression and the orientation with lower expression. Promoters with a strong dependence on the orientation of their site should have a large such ratio. **(B)** For the 75 TFs from Figure 3B whose consensus sites were inserted in their two possible orientations within the same fixed promoter background, shown (red line) is the fraction of TFs (y-axis) for which the ratio between the expression of their strong and weak orientation is at least some ratio *k*, for all possible values of *k* (x-axis). To assess the number of TFs for which the difference in their ratio is statistically significant, shown (green line) is the same plot for a set of 20 promoters that differ only in their barcode sequence and are thus expected to have similar expression levels. As another estimate (blue line), we computed the mean and standard deviation from these same 20 promoters. We then sampled values from a normal distribution defined by these values and generated the same plot using these samples. We used the two above distributions (green and blue) to compute a P-value for each value of k. Also shown is the number and identity of the TFs that pass a significance test corrected for multiple hypothesis using FDR at a confidence level of 0.05, (green and blue).

Supplemental Figure 6. The effect of binding site affinity and promoter background on expression. **(A)** Shown is the expression of a set of promoters in which we inserted the consensus site for Gcn4 within a fixed promoter background (see promoter illustration) without mutations (consensus, top row); with all possible single basepair mutations from the consensus (single mutations, rows 2-22), with two random mutations at each possible combination of two binding site positions (double mutations, rows 23-64); and with ten random mutations in three binding site positions (triple mutations, rows 65-74). The reported logo[1] for Gcn4 is given at the top. Within each row, positions within the binding site in which mutations were performed are colored according to the basepair to which the position was mutated. The consensus site (top row), the reverse complement of the consensus site (row 13), and the site that exists within the native His3 promoter in the yeast genome (row 22) are indicated. **(B)** The in vitro affinity of a Gcn4 site is correlated with the in vivo expression level of a promoter that contains the site. For 36 Gcn4 sites from (A) for which in vitro affinities were reported[2], shown (red points) is the expression level that our method measured for each site (x-axis) against the expression predicted from the in vitro affinity measured for the site using a simple model (y-axis, see Methods). The resulting R^2 of the correspondence is indicated. Also shown (green pentagrams) is a comparison of our expression measurements and previous in vivo expression measurements for Gcn4 site variants done in a different promoter context[3], for four common variants of Gcn4 sites. Note the lower variation in expression that we measured for lower affinity sites.

Supplemental Figure 7. Measuring the effect of binding site affinity on expression. **(A)** Same as Supplemental Figure 6, but for Fhl1 sites. Shown is the expression of a set of promoters in which we inserted the consensus site for Fhl1 within a fixed promoter context (see promoter illustration) without mutations (consensus, top row); with all possible single basepair mutations from the consensus (single mutations, rows 2-25 rows), with two random mutations at each possible combination of two binding site positions (double mutations, rows 26-81); and with ten random mutations in three binding site positions (triple mutations, rows 82-91). The reported logo[4] for Fhl1 is given at the top. Within each row, positions within the binding site in which mutations were performed are colored according to the basepair to which the position was mutated. **(B)** Same as (A), but for Leu3p sites. **(C)** In vitro affinities of Fhl1 sites are correlated with the in vivo expression level of a promoter that contains the site. For 101 Fhl1 sites from (A), shown (red points) is the expression level that our method measured for each site (y-axis) against the score of the site using their published[4] position specific scoring matrices (PSSMs, x-axis). The resulting correlation of the correspondence is indicated. **(D)** Same as (C), but for 117 Leu3p sites from (B).

Supplemental Figure 8. Small changes in binding site location have large effects on expression. **(A)** Same as Figure 3E, shown are the expression levels of promoters in which we inserted the consensus site for Gcn4 at different locations (in 3-4bp increments) within two fixed promoter backgrounds (red and blue lines, backgrounds differ by the presence of a poly(dA:dT) tract). Points correspond to the location in the promoter of the rightmost basepair of the Gcn4 site. For comparison, shown are the expression levels of the original promoter with no Gcn4 sites (black line) and of promoters (gray) in which random mutations of 3bp each time were performed across the non-poly(dA:dT) regions of the promoter, indicating that the effect of changing the location of Gcn4 sites is not due to removal of the original promoter sequence. **(B)** Same as (A), but for Gal4 sites. **(C)** Same as (A), but for Leu3p sites. **(D)** Same as (C), i.e., Leu3p sites in two promoter backgrounds that differ in the location of a poly(dA:dT) tract but where the original background is different than that used in (C).

Supplemental Figure 9. Small changes in the location of TF binding sites have major effects on expression. **(A)** For 13 TFs for which we changed the location of their sites at 7bp increments, shown are comparisons of pairs of promoters in which the location of a TF binding site was modified by *k* basepairs for *k*=7,14,...,84. For every value of *k*, each gray dot corresponds to the (log) ratio between a promoter that contains the TF site at some location (distal site, numerator) and that of the same promoter in which the TF site is located *k* basepairs closer to the core promoter (proximal site, denominator). In addition, for every value of *k* also shown for the individual points is their median (red line), standard error (orange bar), and standard deviation (blue bar). The number of promoter pairs being compared at each value of *k* is indicated and the identity of the TFs is given within a table. (B) Same as (A), in a different promoter background. **(C)** Same as (A), for sites of Gcn4, Gal4, and Leu3p whose locations we changed in 1-4bp increments.

Supplemental Figure 10. The large effects of small changes in the location of TF sites on expression are not due to noise in our measurements. **(A)** For all 1114 promoters that were part of the set of promoters for 17 different TFs whose site location was changed within different promoter backgrounds, shown is the expression that we measured for them in two independent replicates in which we employed different cell sorting strategies. The plot is a subset of the points shown in Figure 1B. This high correspondence between the expression levels of the above promoters between two independent biological replicates (R^2=0.99) demonstrates that the large effects on expression of small basepair changes in TF site locations are not due to noise in our experiment. **(B)** For one specific promoter set in which site locations for Gcn4 were changed in 3-4bp increments, shown are the expression levels that we measured in the two replicates (green and red lines). **(C)** Same as (B), but for a specific promoter set in which Gal4 site locations were changed in 3-4bp increments.

Supplemental Figure 11. The large effects of small changes in the location of TF sites on expression are not due to the different barcode sequences that are unique to each promoter. **(A)** Shown is the correlation between the expression levels of different promoter sets, where in each promoter set Gal4 site locations were changed in 3-4bp increments. Each promoter set corresponds to a different sequence background in which the Gal4 sites were inserted, where the top 7 promoter sets used backgrounds that differ only in the location of a poly(dA:dT) tract, and similarly, the bottom 7 promoter sets used backgrounds that were different from that of the top 7 promoters but within themselves, differed only in the location of a poly(dA:dT) tract. Since each promoter in our library has a unique barcode, the high correlation found between the expression levels of the top 7 promoter sets and between the expression levels of the bottom 7 promoter sets indicates that the jagged expression functions that we measure as a function of TF site location are not due to the different barcode sequences that are unique to each promoter. Note the lower yet still positive correlations between promoter sets of different backgrounds (correlations between promoter sets from the top 7 rows and promoter sets from the bottom 7 rows). **(B)** Same as (A), but for Gcn4 sites. The high correlation between promoter sets with the same background but different locations of poly(dA:dT) tracts can also be observed here. **(C)** Same as (A), but for Leu3p sites. Here the correlations are significantly lower than in (A) and (B).

Supplemental Figure 12. The large effects of small changes in the location of TF sites on expression are not due to removal of the sequences of the original promoter that are replaced when TF sites are inserted. Shown is the correlation between the expression levels of different promoter sets, where in each promoter set locations of a certain TF site was changed. Correlations are shown between promoter sets that differ in both the TF site and promoter background in which the sites were inserted (left column); between promoter sets that differ in the TF site but not in the promoter background (middle column); and between promoter sets that used the same TF site but different promoter backgrounds (right column). In each column, shown are the individual correlations between promoter sets (gray points), as well as the median (red line), standard error (orange bar), and standard deviation (blue bar) of the correlations. Note that the correlation between the promoter sets (jagged functions of expression as a function of site location) of different TFs in the same promoter background is low, highly variable, and sometimes even anti-correlated (middle column), indicating that the jagged functions are not just the result of removal of the original sequences in the promoter that are replaced when the TF sites are inserted.

Supplemental Figure 13. The effect of a repressor site does not show a clear trend when its location is fixed and that of an activator site is changed. **(A)** For the Matα2p-Mcm1p repressor complex, shown are four sets of promoters in which the location of its site within the promoter was fixed, and the location of a Gal4 site was changed, where the four sets differ by the promoter background and presence of a poly(dA:dT) tract. Expression is shown as a ratio between the promoter that contains the repressor site and the same promoter but without the repressor site. **(B)** Same as (A), but when Gcn4 site locations were changed.

Supplemental Figure 14. Expression and the location of a Gcn4 site are related by a ~10bp periodic function. For 4 promoter sets in which we modified locations of Gcn4 sites in 3-4bp increments within promoter backgrounds that differed in the location of a poly(dA:dT) tract, shown is the average of the auto-correlation function of each promoter set when shifting the vector of expression as a function of site location of each promoter set. Note the peaks of the auto-correlation function at 9bp, 21bp, and ~31bp and its valleys at 5bp, 16bp, and 25-26bp. P-values of the auto-correlation are given at significant peaks and valleys.

Supplemental Figure 15. A thermodynamic model for predicting expression from DNA sequence. **(A)** Shown is the ability of four different models to predict the expression of a selected coherent subet of our promoters (all 192 promoters with 0-2 Gcn4 binding sites that were inserted into two different promoter backgrounds) using only their DNA sequence. We devised four different models, described below, for predicting expression from DNA sequence that differ in the components that they each integrate. Predictions were generated using a cross validation scheme, whereby the promoters were randomly partitioned into five non-overlapping sets, and the expression of promoters within each group were predicted using a model whose parameters were learned from the expression of the promoters of the other four sets. Each model is a variant of the model described in Raveh-Sadka et al.[5] in which the binding of each TF is limited according to specific positions within the promoters where we designed its sites. The four different models that we employed are: (1) 'Nucleosomes', in which TF concentrations are set to zero. This model only models the accessibility of the TATA box to Pol II binding; (2) 'Gcn4', in which the nucleosome concentration is set to zero. This model only models the effect of TF binding as in Shea & Ackers[6]; (3) 'Gcn4 and nucleosomes', in which both nucleosomes and TF binding are modeled as in Raveh-Sadka et al.[5]; and (4) 'Gcn4 and nucleosomes with helical repeat', in which we model both nucleosome and TF binding as in Raveh-Sadka et al.[5] but where the interaction weight between Gcn4 and Pol II depends on the helical phasing of their distance. Formally, we define the interaction weight of site i with Pol II as: $w_{i,polII} = w'_{i,polII} \cdot (1 + \varphi sin(\left|\frac{distance}{R}\right| + s))$ where $w'_{i,polII}$ represent the interaction strength, $distance$ is the distance between site I and Pol II (in this work – the TATA box), R is the helical repeat length, S is a helical shift, and φ represents the relative contribution of the helical phasing to the interaction energy. **(B)** Dot-plot of the model predictions on held out test promoters (y-axis) against the expression measured for each promoter (x-axis), for the 'Gcn4' model. **(C)** Same as (B), but for the 'Gcn4 and nucleosomes' model. **(D)** Same as (B), but for the 'Gcn4 and nucleosomes with helical repeat' model.

Supplemental Figure 16. The stimulatory effect of poly(dA:dT) tracts increases with their length. For promoter backgrounds that differ in the presence of sites for Gal4 and Gcn4 with different strengths and a poly(dA:dT) tract close to the core promoter, shown are expression levels of promoters in which poly(dA:dT) tracts of varying lengths were inserted at the same location. Note that, on average, longer poly(dA:dT) tracts result in higher expression levels across all promoter sets.

Supplemental Figure 17. Assessing the relative contribution of number of sites. We fit different logistic functions to the expression levels of the promoters from Figure 5B. in which we inserted the consensus site for Gal4 in all 2^5=32 possible combinations of sites at five predefined locations within the promoter. The leftmost column corresponds to a logistic function in which sites at all locations have the same weight (parameter). In the next column, each site has a different weight. Each subsequent column corresponds to the same logistic function as in the previous column with the addition of another weight for the presence of a specific pair of sites indicated in the x-axis. Site pair weights are ordered by the weight that when added to the previous logistic function, provides the largest improvement in the fit to the data. Note that in both promoter backgrounds, weights for neighboring site locations are added first. For each logistic function, shown is the fraction of the variance explained by it (y-axis). For comparison, shown are the same fits to a dataset in which we permuted the expression levels of promoters that have the same number of Gal4 sites. The worse fits observed indicate a coherent contribution for sites at different locations. **(B)** Same as (A) but when fitting logistic functions to promoters of Gcn4, in which all 2^7=128 possible combinations of sites were inserted at seven predefined locations within the promoter. For both (A) and (B), the improvement in the fit when sites at all locations have the same weight (leftmost column, R^2=0.66-0.89) to that in which there is a different weight for each site location (second column from left, R^2=0.78-0.91) suggests that sites have different contributions to expression at different promoter locations. Since it has more parameters, the improved fit of this latter model to the data is expected, but the amount by which it better fits provides an upper bound for the location-dependent contribution of sites to expression, at least for this logistic function formulation. Similarly, models with weights for specific pairs assess the expression contribution that pairs of sites at specific locations have beyond their individual contributions. Notably, (A) above (Gal4), adding just one more parameter representing a negative weight between the pair of sites at the two locations closest to the gene start was sufficient for fitting 98% of the expression variability in both contexts. Since the ends of these sites are only one basepair apart (compared to a 5bp difference between all other adjacent Gal4 sites), this result suggests that two Gal4 molecules sterically occlude each other at such distances (see also **Fig. S19A**). Although not as striking, (B) shows a similar behavior for Gcn4, whereby the first five parameters that provide the largest improvement when added to the model correspond to weights for pairs of sites that are adjacent, and in both contexts, weights for these pairs are lower than those for pairs of non-adjacent sites (see also **Fig. S19B**). Since ends of adjacent Gcn4 sites were always separated by 5bp, these results suggest that Gcn4 sites that are 5bp apart may partially occlude each other, although to a much lesser extent than Gal4 sites that are one basepair apart.

Supplemental Figure 18. Assessing the relative contribution of number of sites and site locations to expression. (A) For one of the promoter background in which we inserted the consensus site for Gal4 in all 2^5=32 possible combinations of sites at five predefined locations within the promoter, shown are fits of three different logistic functions to all 32 expression measurements. In the leftmost logistic function all five site locations have the same weight parameter. In the middle logistic function, there is a separate weight (logistic parameter) for each site. In the rightmost logistic function, there are separate weights for each site as well as weights for the different pairs of sites. Each fit shows the fits of the logistic function (y-axis) against the measured expression of each promoter (x-axis) with colors representing the number of Gal4 sites in the corresponding promoter. In the rightmost logistic function, the inset heatmap shows the values fitted for the different site pair weights. Note the lower values fit to pairs of sites that are adjacent in their promoter location. **(B)** Same as (A), for the other promoter background in which Gal4 sites were inserted. **(C-D)** Same as (A) and (B) but for Gcn4, for which all 2^7=128 possible combinations of sites were inserted at seven predefined locations within the promoter.

Supplemental Figure 19. Possible steric hindrance between closely spaced Gal4 and Gcn4 sites. (A) We selected six promoter backgrounds that differed in the location of a poly(dA:dT) tract and of a single Gal4 site. In each promoter background, we then inserted another Gal4 site at varying distances from the already existing Gal4 site. For each promoter background and at every tested distance between the two Gal4 sites, shown is the (log) ratio between the expression of a promoter that contains the two Gal4 sites at locations *i* and *j* and the maximum between the expression of the promoter that contains a single Gal4 site at location *i* and the promoter that contains a single Gal4 site at location *j*. Note that in most cases, the expression of a promoter that contains Gal4 sites whose ends are 1bp apart is not higher than the maximum of the two promoters that contain the individual sites, indicating a possible steric hindrance between the two Gal4 sites at this distance. **(B)** Same as (A), but for Gcn4 sites, for which the shortest distance tested between the ends of neighboring sites was 5bp. Here too, the lower ratios observed at shorter distances between site pairs is suggestive of a possible steric hindrance between Gcn4 sites at this distance, although the effect is significantly smaller than that observed for Gal4 sites in (A).

Supplemental Figure 20. The contribution of a TF site diminishes with the number of sites present in the promoter to which it is added. (A) For one promoter background, shown is the (log) ratio between a promoter in which the *k*-th site was added and the same promoter without that *k*-th site, for *k=1,2,…,7,* grouped at the x-axis by the value of *k*. At each value *k,* the individual ratios are shown (gray points), as well as their median (red line), standard error (orange bar), and standard deviation (blue bar). **(B)** Same as (A), for another promoter background.

Supplemental Figure 21. Expression level is a non-monotonic function of the number of Rgt1p sites. **(A)** We inserted up to seven (top two rows) or five (bottom row) Rgt1p sites in increments of one site within two different promoter backgrounds with no known TF sites (top row, green and red bars), two different promoter backgrounds that contain a single Leu3p site (middle row, green and red bars), and two different promoter backgrounds that contain a single Gal4 site (bottom row, green and red bars). Bars show the expression levels measured for these promoters, showing that in most cases expression increases with the addition of each of the first 3 Rgt1p sites but then decreases greatly with the addition of the fourth or fifth Rgtp1 site, where the expression is much lower than that of the promoter without any Rgt1p sites. **(B)** For the repressor complex Matα2p-Mcm1p, shown is the expression of a promoter that contains zero, one, or two sites in two different backgrounds (green and red bars), indicating a repressive effect for Matα2p-Mcm1p that becomes greater with the addition of a second Matα2p-Mcm1p site.

References

1 MacIsaac, K. D. *et al.* An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7**, 113 (2006).

2 Nutiu, R. *et al.* Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat Biotechnol* **29**, 659-664 (2011).

3 Hill, D. E., Hope, I. A., Macke, J. P. & Struhl, K. Saturation mutagenesis of the yeast his3 regulatory site: requirements for transcriptional induction and for binding by GCN4 activator protein. *Science* **234**, 451-457 (1986).

4 Zhu, C. *et al.* High-resolution DNA binding specificity analysis of yeast transcription factors. *Genome Res* **19**, 556-566 (2009).

5 Raveh-Sadka, T., Levo, M. & Segal, E. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res* **19**, 1480-1496 (2009).

6 Shea, M. A. & Ackers, G. K. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Mol Biol* **181**, 211-230, doi:0022-2836(85)90086-5 [pii] (1985).

Supplementary figures

Figure S1

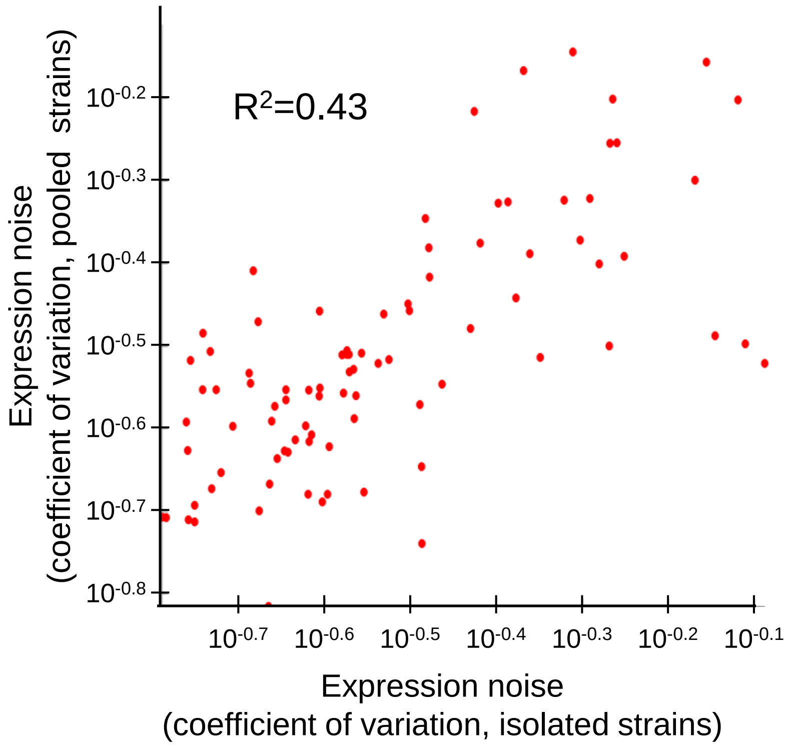

R²=0.43
Expression noise
(coefficient of variation, pooled strains)
10^-0.2
10^-0.3
10^-0.4
10^-0.5
10^-0.6
10^-0.7
10^-0.8
10^-0.7
10^-0.6
10^-0.5
10^-0.4
10^-0.3
10^-0.2
10^-0.1
Expression noise
(coefficient of variation, isolated strains)

Figure S2

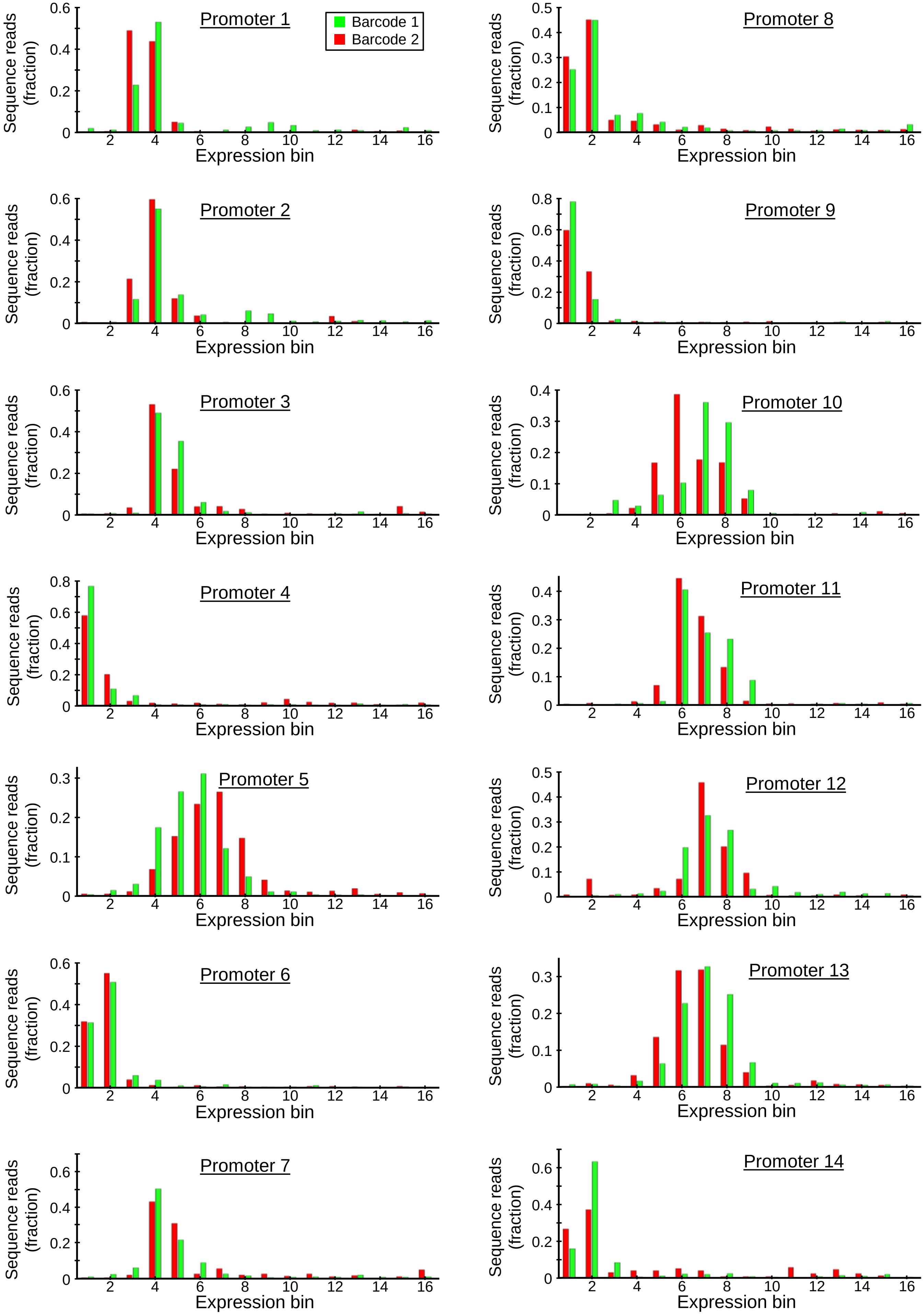

Promoter 1
🟩 Barcode 1
🟥 Barcode 2
Promoter 2
Promoter 3
Promoter 4
Promoter 5
Promoter 6
Promoter 7
Promoter 8
Promoter 9
Promoter 10
Promoter 11
Promoter 12
Promoter 13
Promoter 14
Sequence reads (fraction)
Expression bin
0
0.1
0.2
0.3
0.4
0.5
0.6
0.8
2
4
6
8
10
12
14
16

Figure S3

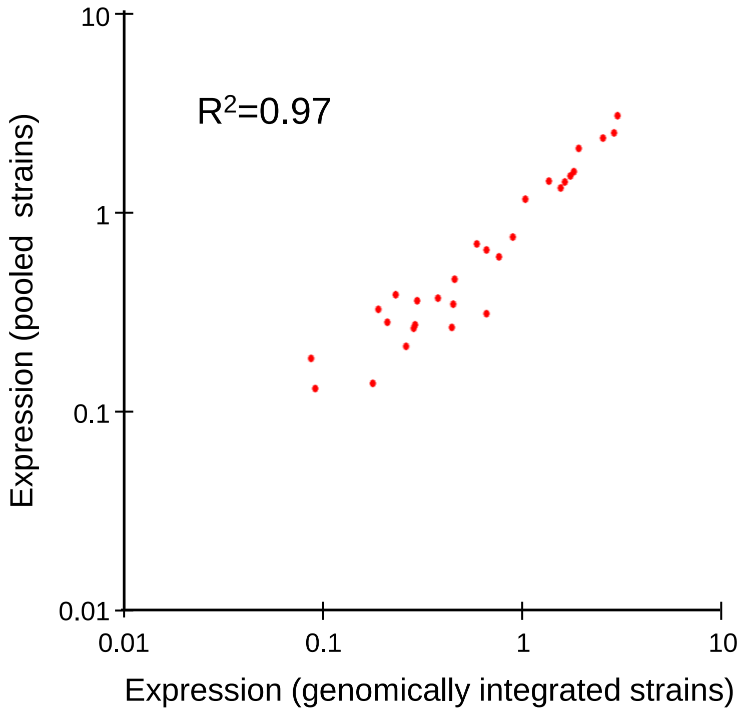

R^2=0.97
10
1
0.1
0.01
Expression (pooled strains)
0.01
0.1
1
10
Expression (genomically integrated strains)

Figure S4

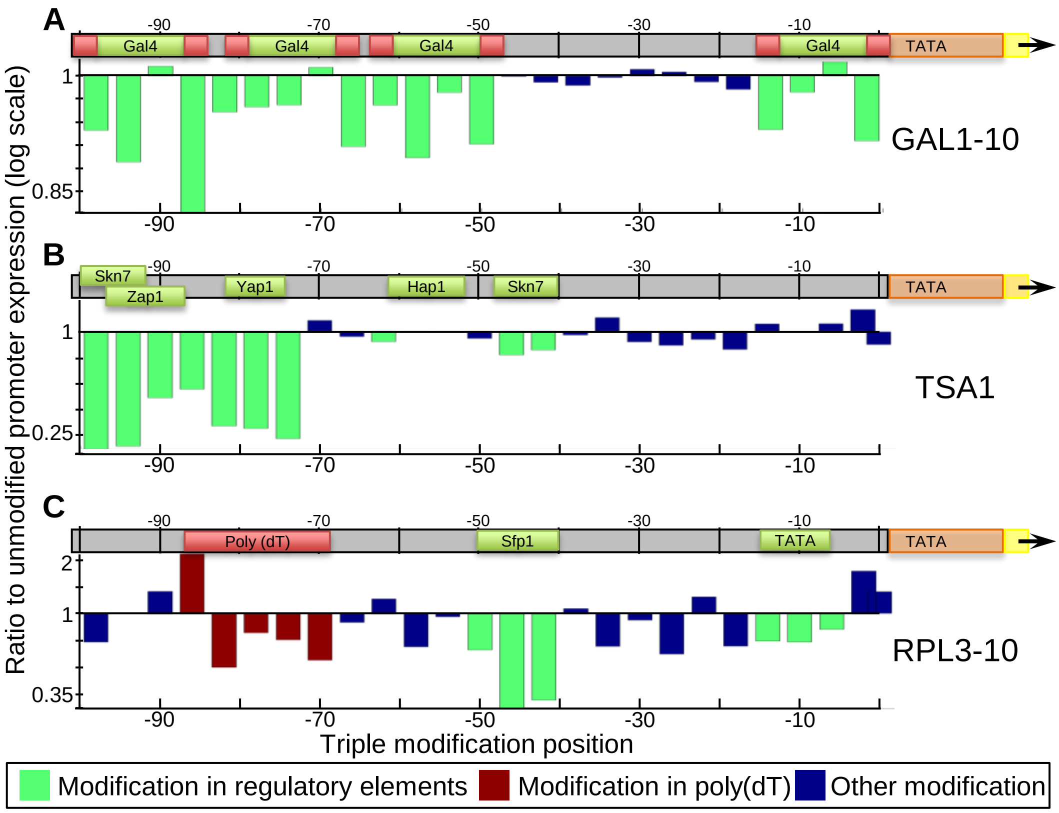

A
-90
-70
-50
-30
-10
Gal4
Gal4
Gal4
Gal4
TATA
1
0.85
GAL1-10
-90
-70
-50
-30
-10
B
Skn7
Zap1
Yap1
Hap1
Skn7
TATA
1
0.25
TSA1
C
Poly (dT)
Sfp1
TATA
2
1
0.35
RPL3-10
Ratio to unmodified promoter expression (log scale)
Triple modification position
Modification in regulatory elements
Modification in poly(dT)
Other modification

Figure S5

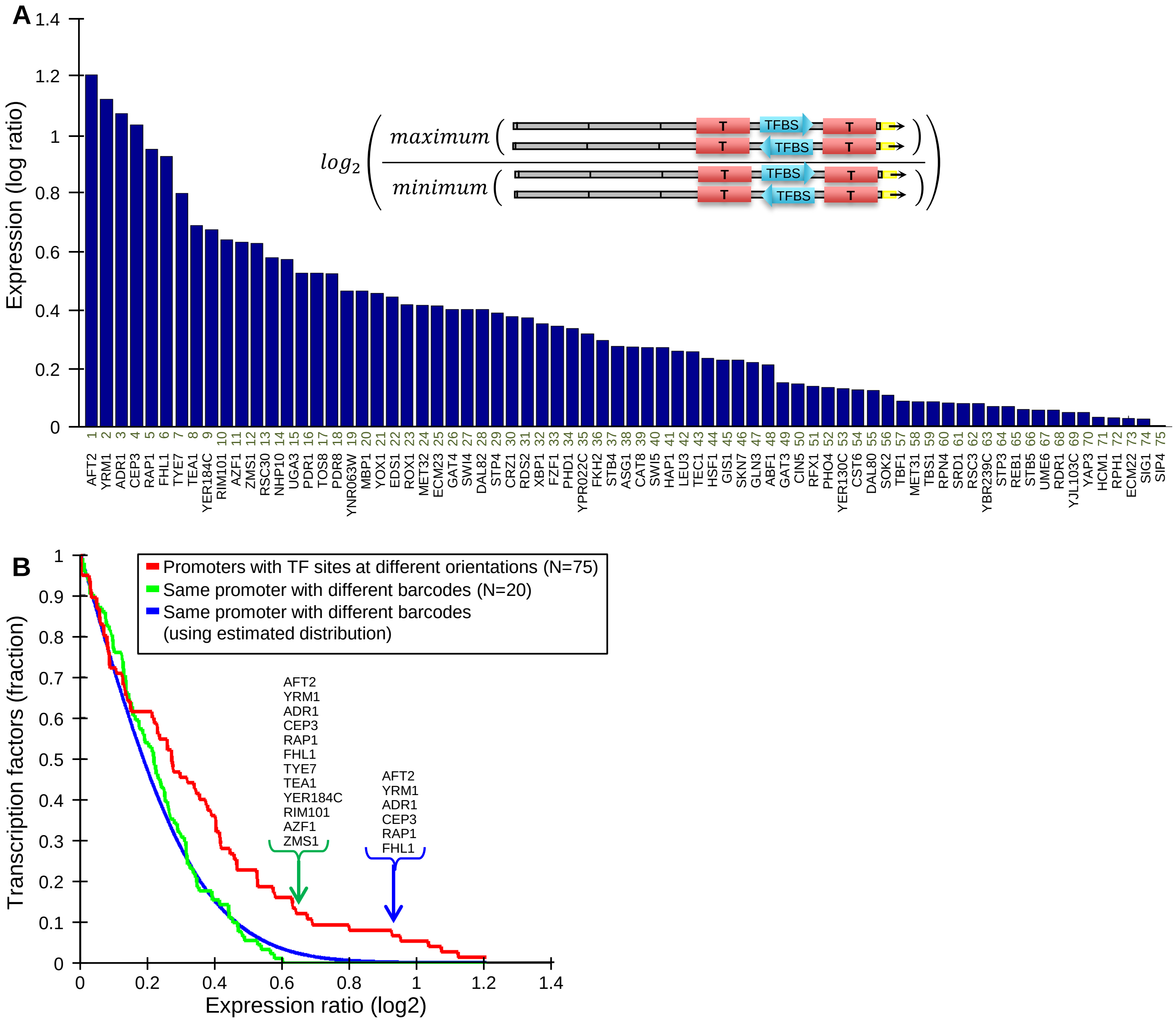

A
Expression (log ratio)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
log_2
maximum
minimum
T
TFBS
1 AFT2
2 YRM1
3 ADR1
4 CEP3
5 RAP1
6 FHL1
7 TYE7
8 TEA1
9 YER184C
10 RIM101
11 AZF1
12 ZMS1
13 RSC30
14 NHP10
15 UGA3
16 PDR1
17 TOS8
18 PDR8
19 YNR063W
20 MBP1
21 YOX1
22 EDS1
23 ROX1
24 MET32
25 ECM23
26 GAT4
27 SWI4
28 DAL82
29 STP4
30 CRZ1
31 RDS2
32 XBP1
33 FZF1
34 PHD1
35 YPR022C
36 FKH2
37 STB4
38 ASG1
39 CAT8
40 SWI5
41 HAP1
42 LEU3
43 TEC1
44 HSF1
45 GIS1
46 SKN7
47 GLN3
48 ABF1
49 GAT3
50 CIN5
51 RFX1
52 PHO4
53 YER130C
54 CST6
55 DAL80
56 SOK2
57 TBF1
58 MET31
59 TBS1
60 RPN4
61 SRD1
62 RSC3
63 YBR239C
64 STP3
65 REB1
66 STB5
67 UME6
68 RDR1
69 YJL103C
70 YAP3
71 HCM1
72 RPH1
73 ECM22
74 SIG1
75 SIP4
B
Transcription factors (fraction)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Expression ratio (log2)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
🟥 Promoters with TF sites at different orientations (N=75)
🟩 Same promoter with different barcodes (N=20)
🟦 Same promoter with different barcodes (using estimated distribution)
AFT2
YRM1
ADR1
CEP3
RAP1
FHL1
TYE7
TEA1
YER184C
RIM101
AZF1
ZMS1
AFT2
YRM1
ADR1
CEP3
RAP1
FHL1

Figure S6

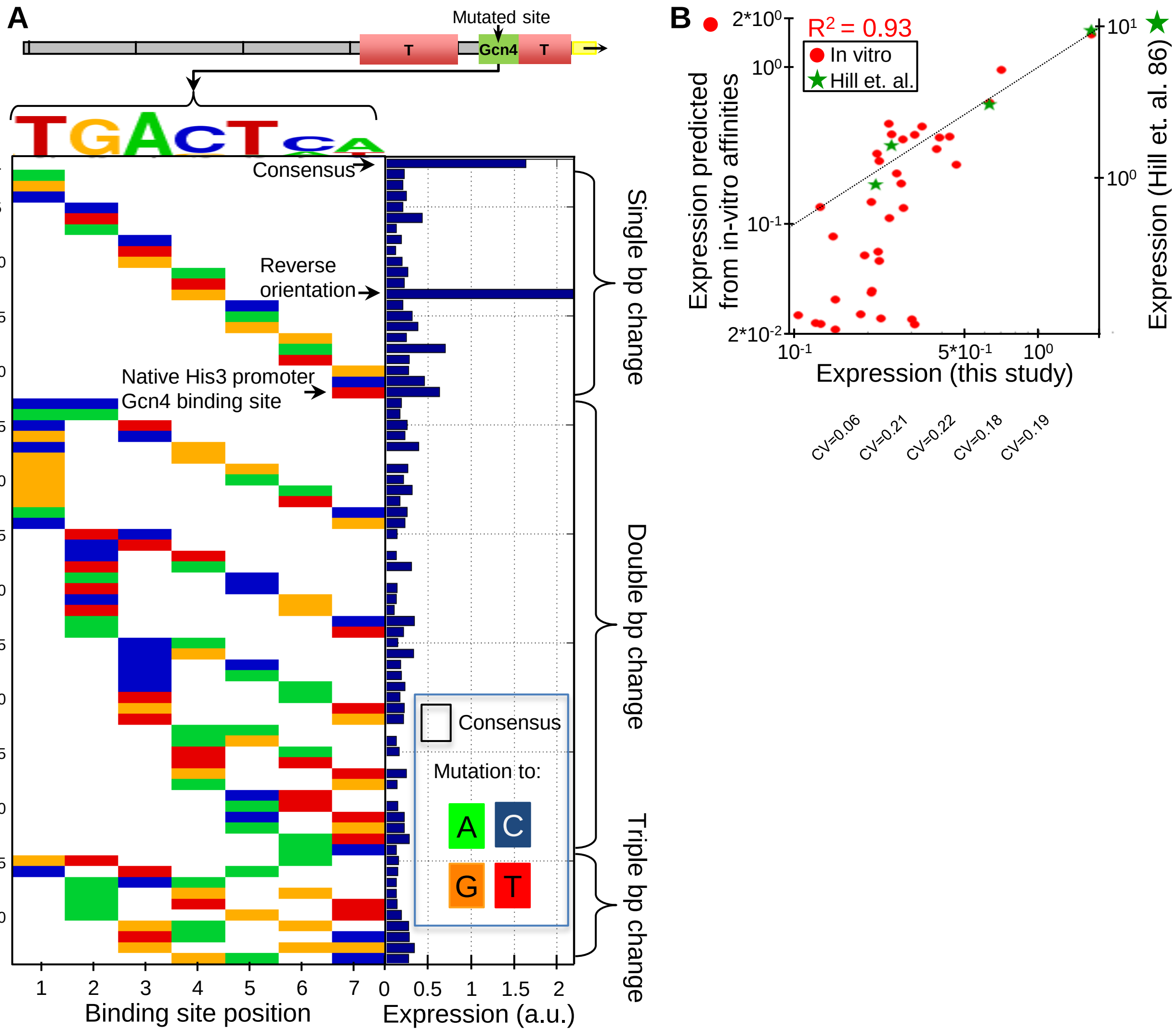

A
Mutated site
T
Gcn4
T
TGACTCA
Consensus
Reverse orientation
Native His3 promoter Gcn4 binding site
1
5
10
15
20
25
30
35
40
45
50
55
60
65
70
1 2 3 4 5 6 7
Binding site position
0 0.5 1 1.5 2
Expression (a.u.)
Consensus
Mutation to:
A
C
G
T
Single bp change
Double bp change
Triple bp change
B
$R^2 = 0.93$
In vitro
Hill et. al.
$2*10^0$
10^0
10^{-1}
$2*10^{-2}$
10^{-1}
$5*10^{-1}$
10^0
10^1
10^0
Expression predicted from in-vitro affinities
Expression (this study)
Expression (Hill et. al. 86)
CV=0.06
CV=0.21
CV=0.22
CV=0.18
CV=0.19

Figure S7

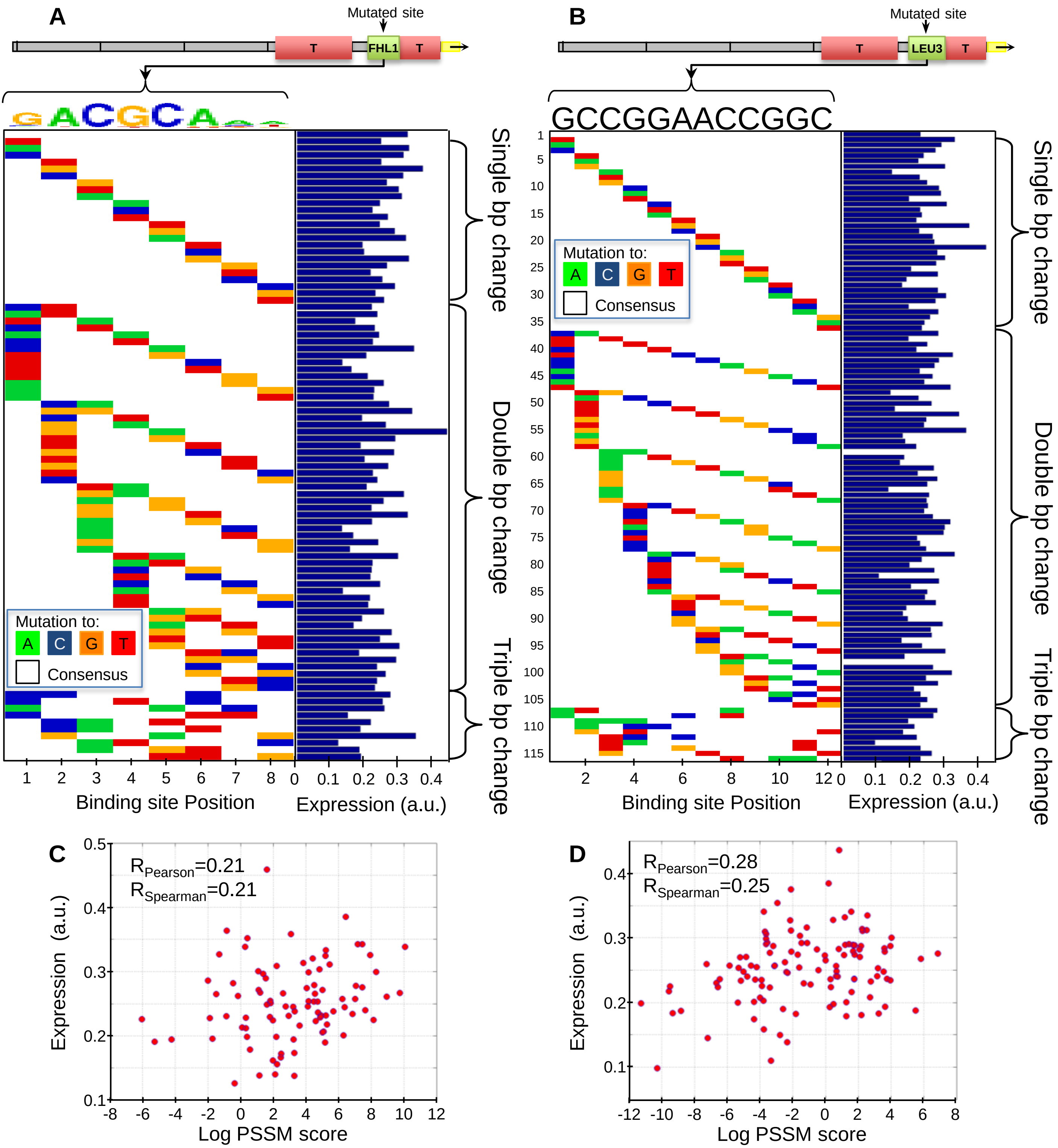

A
Mutated site
T
FHL1
T
GACGCA
Mutation to:
A C G T
Consensus
Binding site Position
Expression (a.u.)
Single bp change
Double bp change
Triple bp change
B
Mutated site
T
LEU3
T
GCCGGAACCGGC
Mutation to:
A C G T
Consensus
Binding site Position
Expression (a.u.)
Single bp change
Double bp change
Triple bp change
C
$R_{Pearson}$=0.21
$R_{Spearman}$=0.21
Expression (a.u.)
Log PSSM score
D
$R_{Pearson}$=0.28
$R_{Spearman}$=0.25
Expression (a.u.)
Log PSSM score

Figure S8

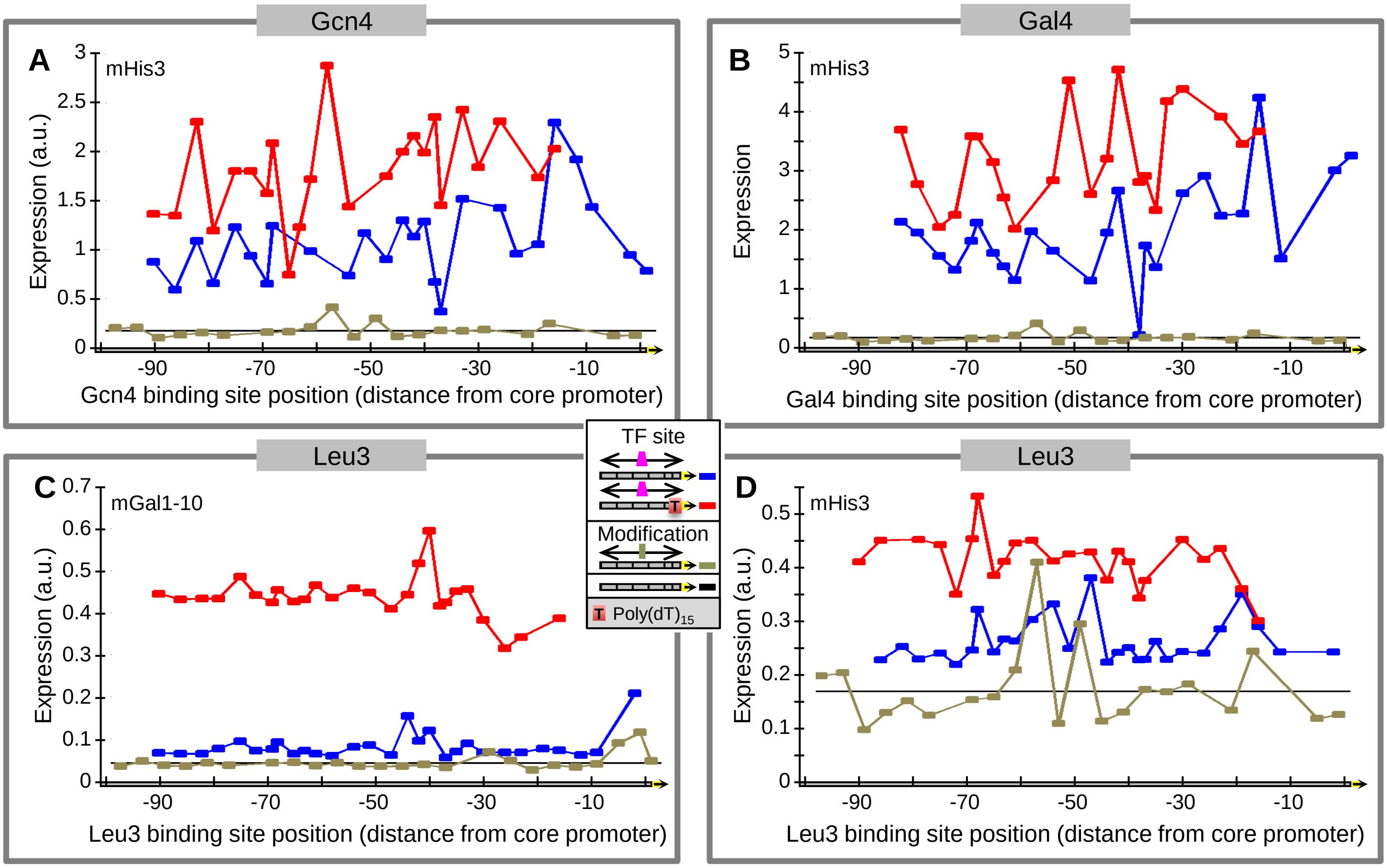
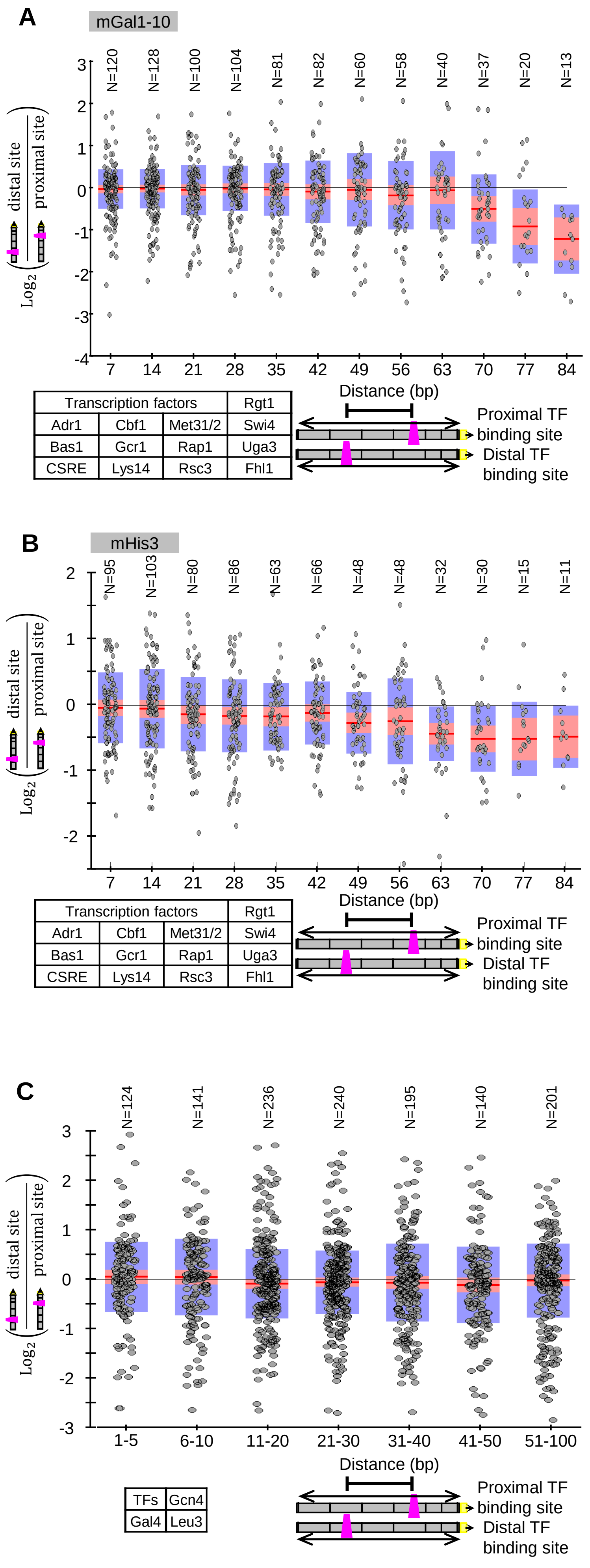

Gcn4
A
mHis3
Expression (a.u.)
3
2.5
2
1.5
1
0.5
0
-90
-70
-50
-30
-10
Gcn4 binding site position (distance from core promoter)
Gal4
B
mHis3
Expression
5
4
3
2
1
0
-90
-70
-50
-30
-10
Gal4 binding site position (distance from core promoter)
TF site
Modification
T
T Poly(dT)$_{15}$
Leu3
C
mGal1-10
Expression (a.u.)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-90
-70
-50
-30
-10
Leu3 binding site position (distance from core promoter)
Leu3
D
mHis3
Expression (a.u.)
0.5
0.4
0.3
0.2
0.1
0
-90
-70
-50
-30
-10
Leu3 binding site position (distance from core promoter)

Figure S9



Transcription factors			Rgt1
Adr1	Cbf1	Met31/2	Swi4
Bas1	Gcr1	Rap1	Uga3
CSRE	Lys14	Rsc3	Fhl1

Transcription factors			Rgt1
Adr1	Cbf1	Met31/2	Swi4
Bas1	Gcr1	Rap1	Uga3
CSRE	Lys14	Rsc3	Fhl1

TFs	Gcn4
Gal4	Leu3

Figure S10

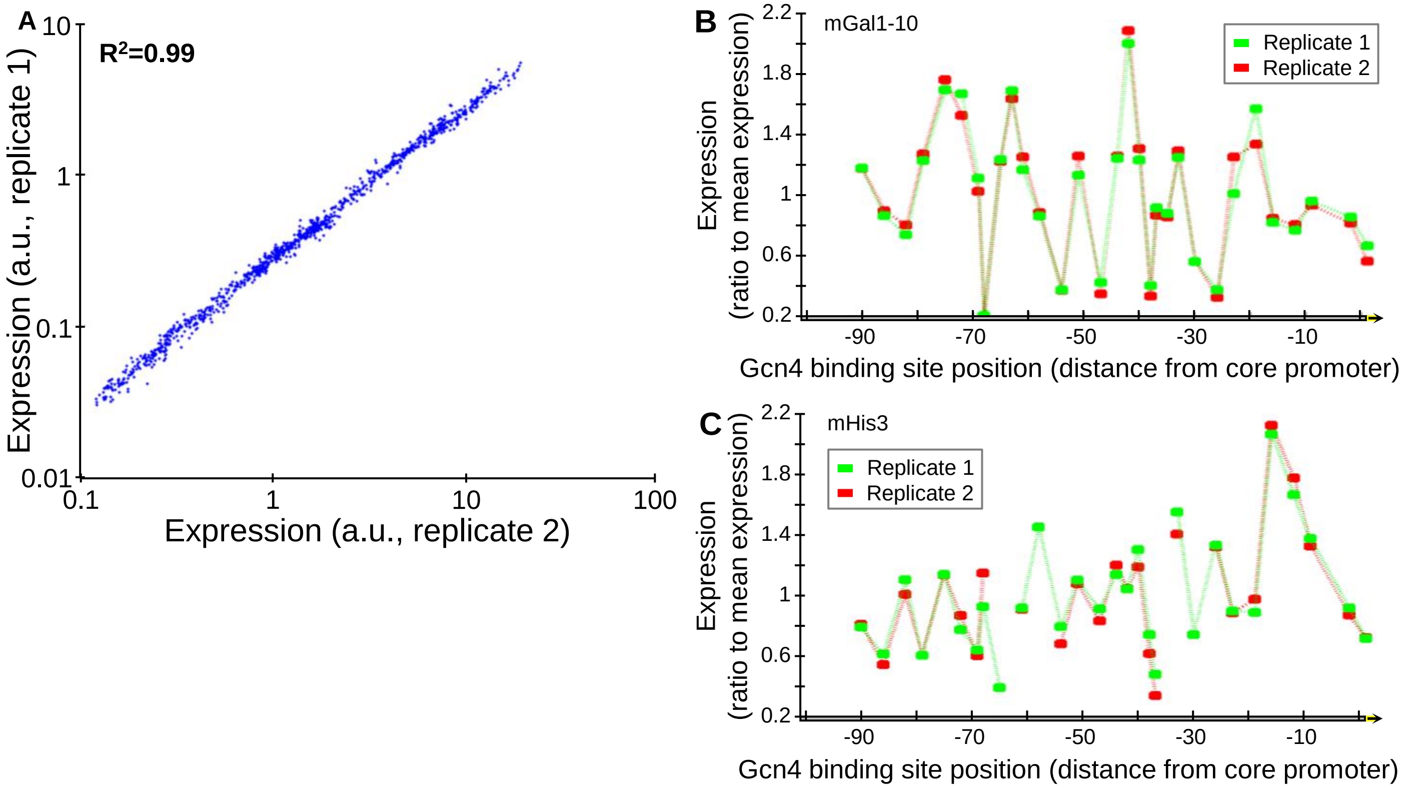

A
R²=0.99
Expression (a.u., replicate 1)
10
1
0.1
0.01
0.1
1
10
100
Expression (a.u., replicate 2)
B
mGal1-10
🟩 Replicate 1
🟥 Replicate 2
Expression
(ratio to mean expression)
2.2
1.8
1.4
1
0.6
0.2
-90
-70
-50
-30
-10
Gcn4 binding site position (distance from core promoter)
C
mHis3
🟩 Replicate 1
🟥 Replicate 2
Expression
(ratio to mean expression)
2.2
1.8
1.4
1
0.6
0.2
-90
-70
-50
-30
-10
Gcn4 binding site position (distance from core promoter)

Figure S11

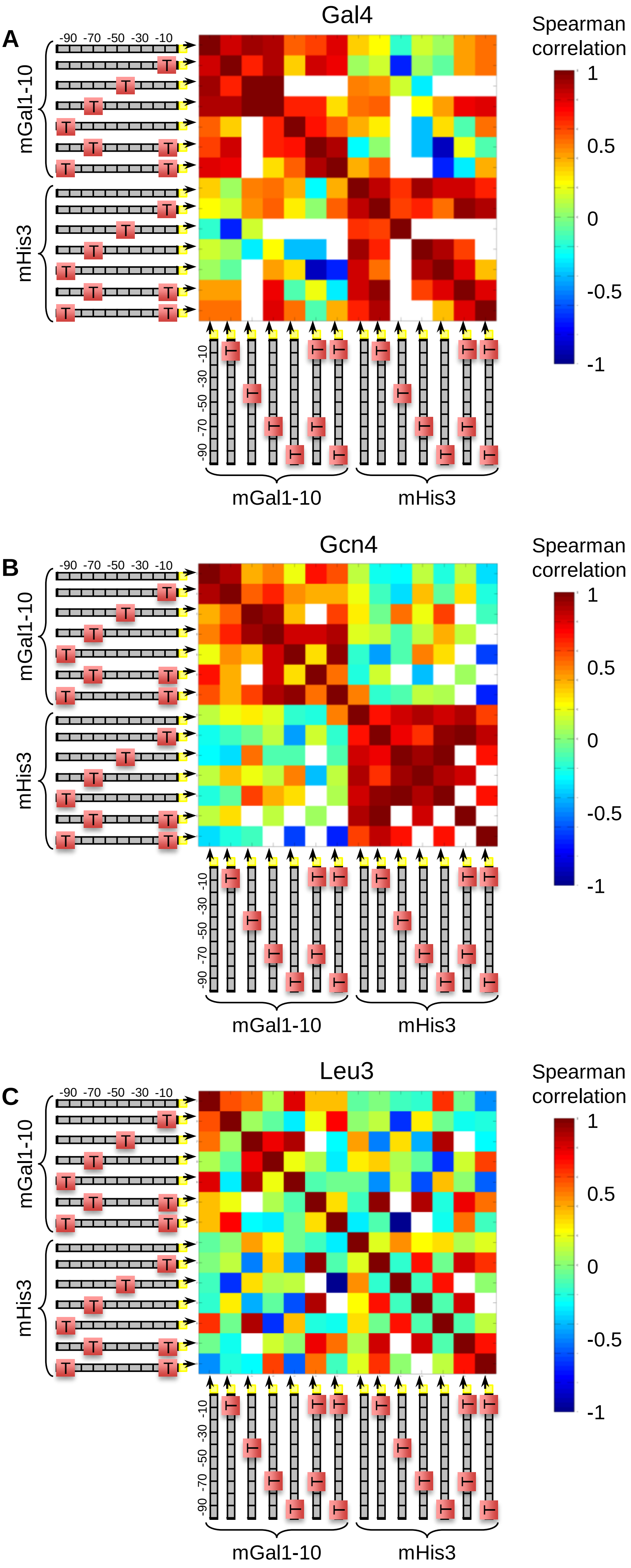

A
Gal4
Spearman correlation
1
0.5
0
-0.5
-1
-90 -70 -50 -30 -10
T
mGal1-10
mHis3
B
Gcn4
Spearman correlation
1
0.5
0
-0.5
-1
-90 -70 -50 -30 -10
T
mGal1-10
mHis3
C
Leu3
Spearman correlation
1
0.5
0
-0.5
-1
-90 -70 -50 -30 -10
T
mGal1-10
mHis3

Figure S12

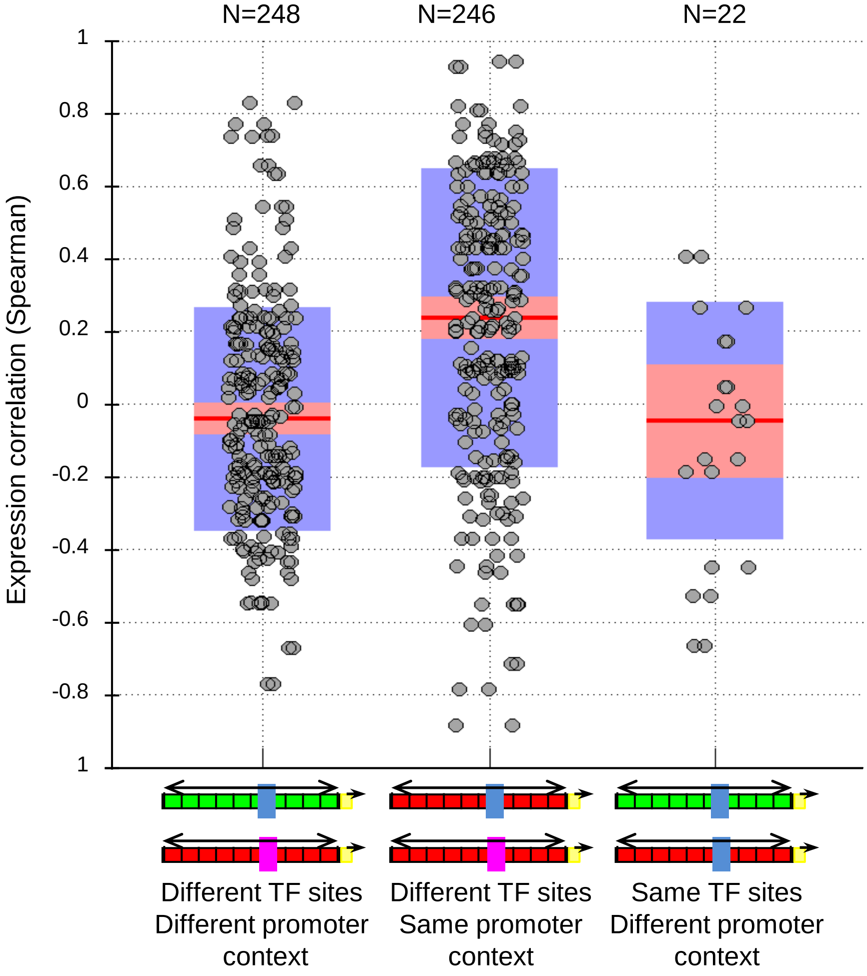

N=248
N=246
N=22
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
1
Expression correlation (Spearman)
Different TF sites
Different promoter
context
Different TF sites
Same promoter
context
Same TF sites
Different promoter
context

Figure S13

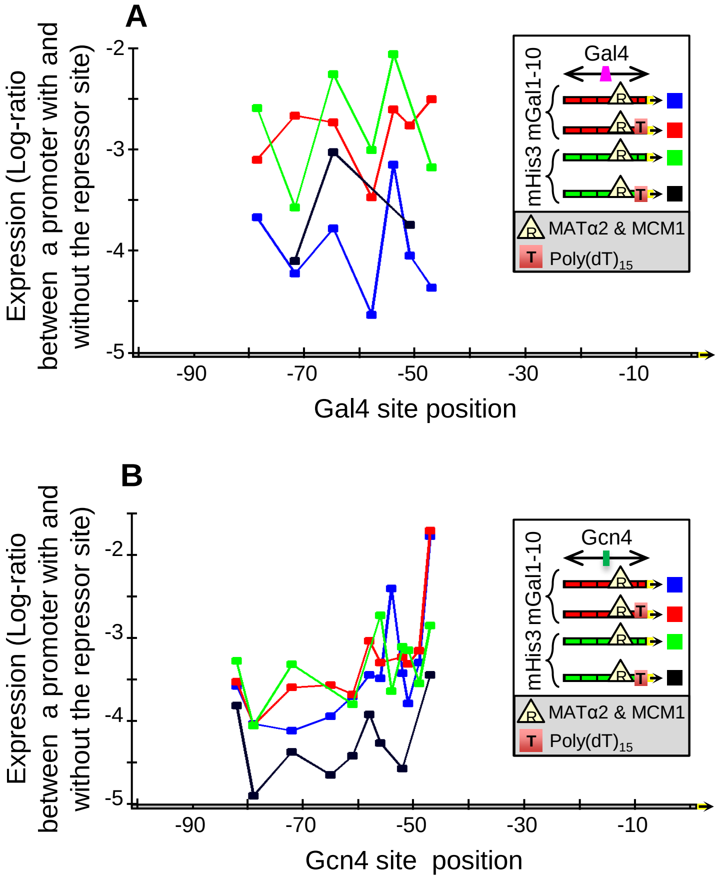

A
Expresion (Log-ratio between a promoter with and without the repressor site)
-2
-3
-4
-5
-90
-70
-50
-30
-10
Gal4 site position
Gal4
mHis3 mGal1-10
R
T
MATα2 & MCM1
Poly(dT)$_{15}$
B
Expresion (Log-ratio between a promoter with and without the repressor site)
-2
-3
-4
-5
-90
-70
-50
-30
-10
Gcn4 site position
Gcn4
mHis3 mGal1-10
R
T
MATα2 & MCM1
Poly(dT)$_{15}$

Figure S14

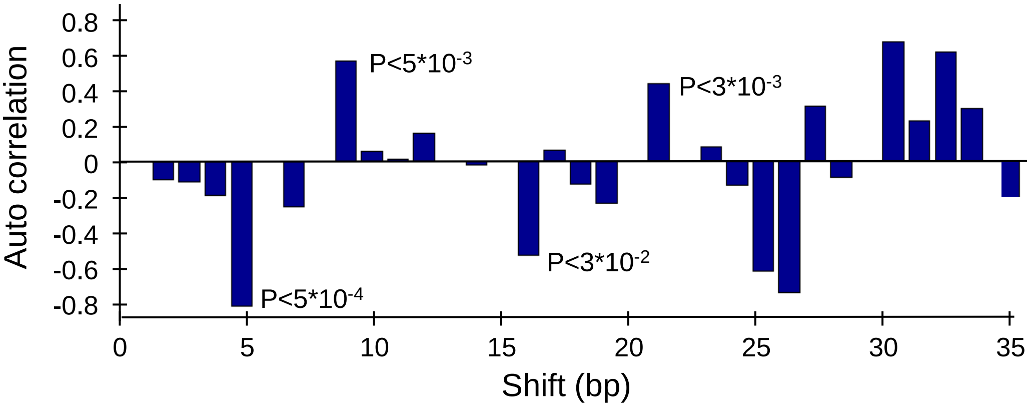

Auto correlation
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
P<5*10-3
P<3*10-3
P<5*10-4
P<3*10-2
0
5
10
15
20
25
30
35
Shift (bp)

Figure S15

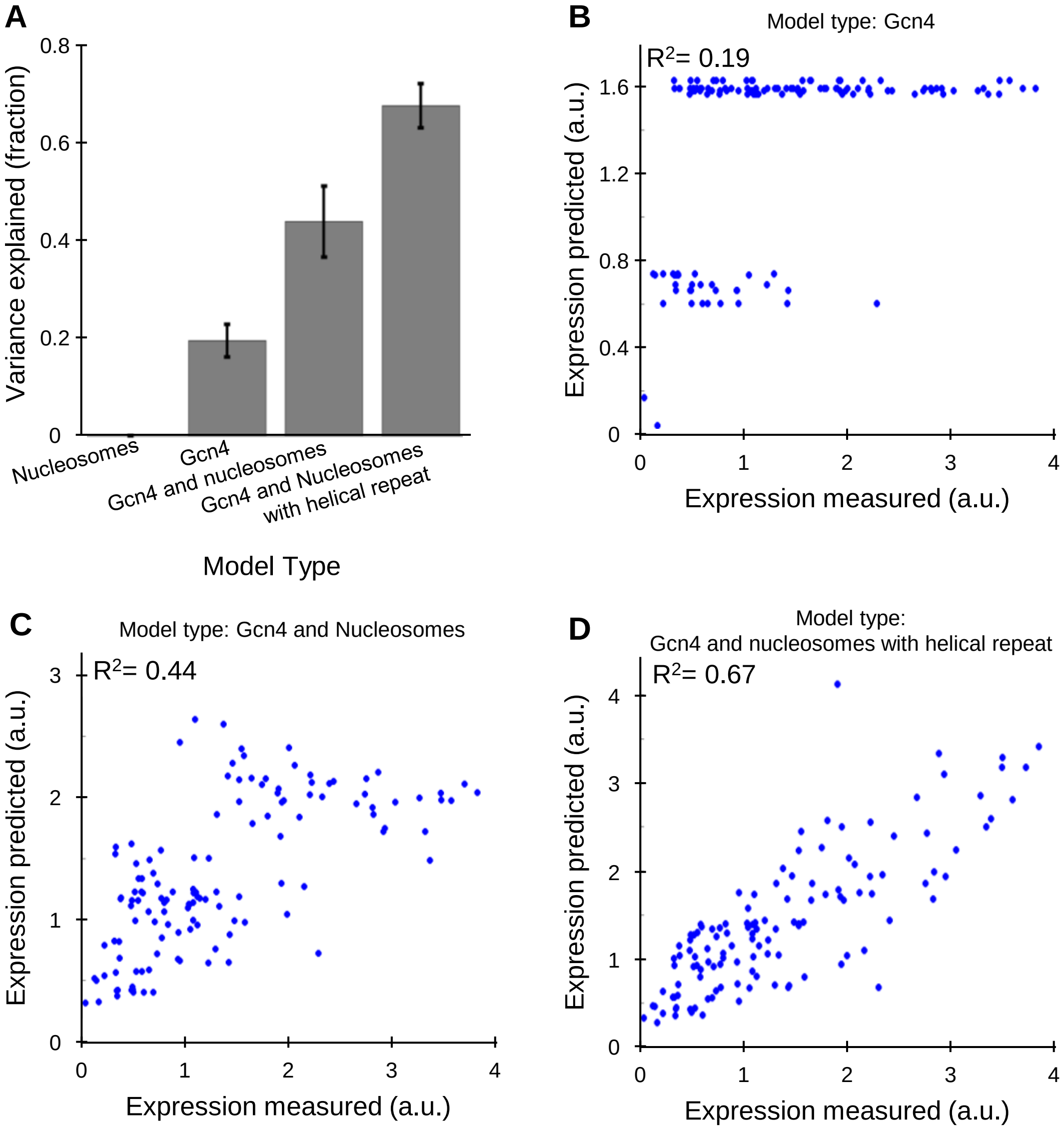

A
Variance explained (fraction)
0.8
0.6
0.4
0.2
0
Nucleosomes
Gcn4
Gcn4 and nucleosomes
Gcn4 and Nucleosomes with helical repeat
Model Type
B
Model type: Gcn4
R^2= 0.19
Expression predicted (a.u.)
1.6
1.2
0.8
0.4
0
0
1
2
3
4
Expression measured (a.u.)
C
Model type: Gcn4 and Nucleosomes
R^2= 0.44
Expression predicted (a.u.)
3
2
1
0
0
1
2
3
4
Expression measured (a.u.)
D
Model type:
Gcn4 and nucleosomes with helical repeat
R^2= 0.67
Expression predicted (a.u.)
4
3
2
1
0
0
1
2
3
4
Expression measured (a.u.)

Figure S16

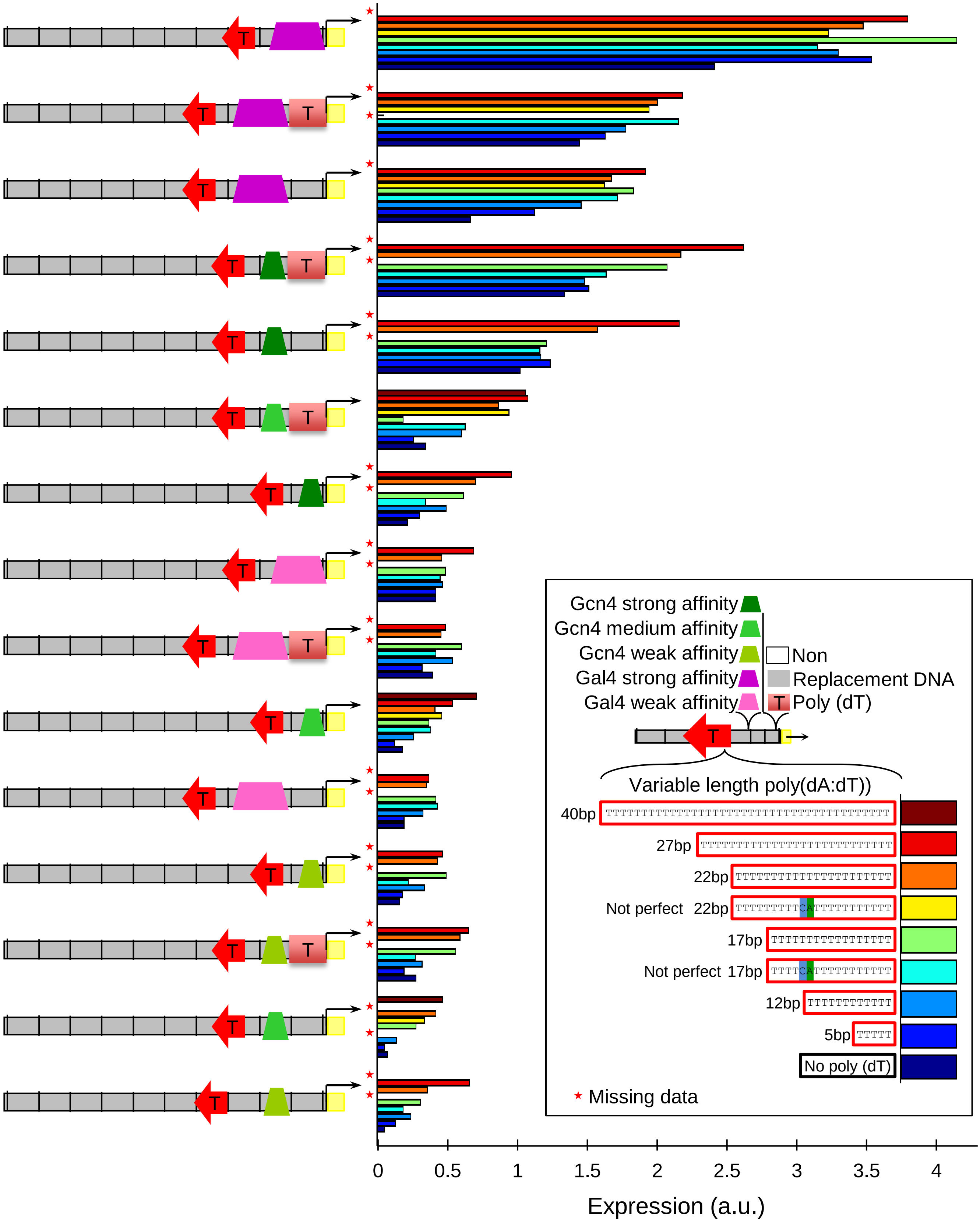
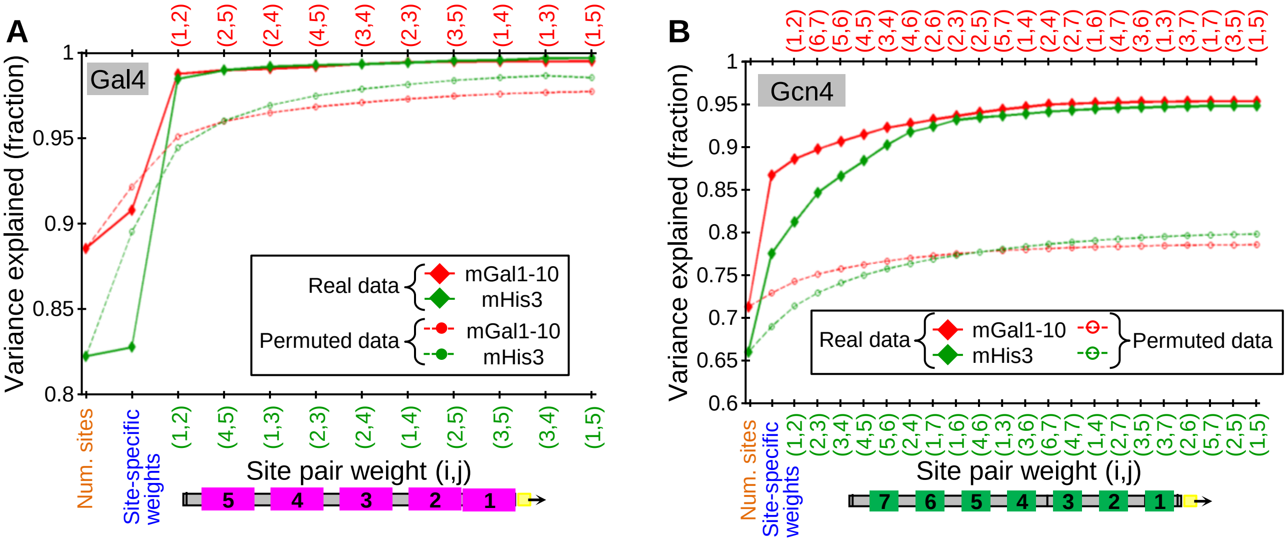

T
Gcn4 strong affinity
Gcn4 medium affinity
Gcn4 weak affinity
Gal4 strong affinity
Gal4 weak affinity
Non
Replacement DNA
Poly (dT)
Variable length poly(dA:dT))
40bp TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
27bp TTTTTTTTTTTTTTTTTTTTTTTTTTT
22bp TTTTTTTTTTTTTTTTTTTTTT
Not perfect 22bp TTTTTTTTTCATTTTTTTTTTT
17bp TTTTTTTTTTTTTTTTT
Not perfect 17bp TTTTCATTTTTTTTTTT
12bp TTTTTTTTTTTT
5bp TTTTT
No poly (dT)
★ Missing data
0
0.5
1
1.5
2
2.5
3
3.5
4
Expression (a.u.)

Figure S17



A
Gal4
Variance explained (fraction)
1
0.95
0.9
0.85
0.8
(1,2) (2,5) (2,4) (4,5) (3,4) (2,3) (3,5) (1,4) (1,3) (1,5)
Num. sites
Site-specific weights
(1,2) (4,5) (1,3) (2,3) (2,4) (1,4) (2,5) (3,5) (3,4) (1,5)
Site pair weight (i,j)
Real data
mGal1-10
mHis3
Permuted data
mGal1-10
mHis3
5
4
3
2
1
B
Gcn4
Variance explained (fraction)
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
(1,2) (6,7) (5,6) (4,5) (3,4) (4,6) (2,6) (2,3) (2,5) (5,7) (1,4) (2,4) (2,7) (1,6) (4,7) (3,6) (1,3) (3,7) (1,7) (3,5) (1,5)
Num. sites
Site-specific weights
(1,2) (2,3) (3,4) (4,5) (5,6) (2,4) (1,7) (1,6) (4,6) (1,3) (3,6) (6,7) (4,7) (1,4) (2,7) (3,5) (3,7) (2,6) (5,7) (2,5) (1,5)
Site pair weight (i,j)
Real data
mGal1-10
mHis3
Permuted data
7
6
5
4
3
2
1

Figure S18

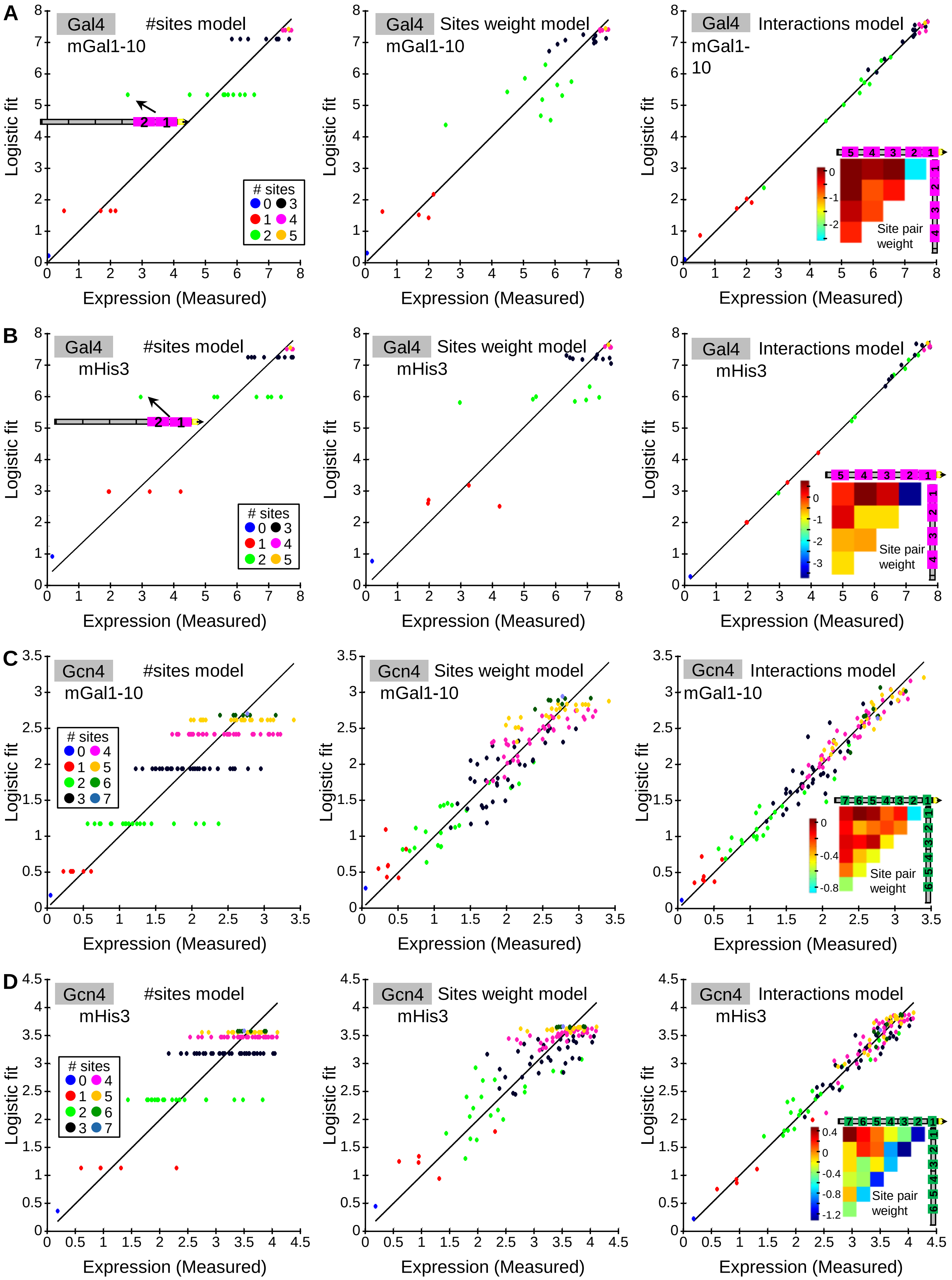

A
Gal4
#sites model
mGal1-10
Sites weight model
Interactions model
mGal1-
10
Logistic fit
Expression (Measured)
# sites
0 3
1 4
2 5
2 1
5 4 3 2 1
1 2 3 4
0
-1
-2
Site pair
weight
B
Gal4
#sites model
mHis3
Sites weight model
Interactions model
0
-1
-2
-3
C
Gcn4
#sites model
mGal1-10
Sites weight model
Interactions model
# sites
0 4
1 5
2 6
3 7
7 6 5 4 3 2 1
1 2 3 4 5 6
0
-0.4
-0.8
D
Gcn4
#sites model
mHis3
Sites weight model
Interactions model
0.4
0
-0.4
-0.8
-1.2
0
1
2
3
4
5
6
7
8
0.5
1.5
2.5
3.5
4.5

Figure S19

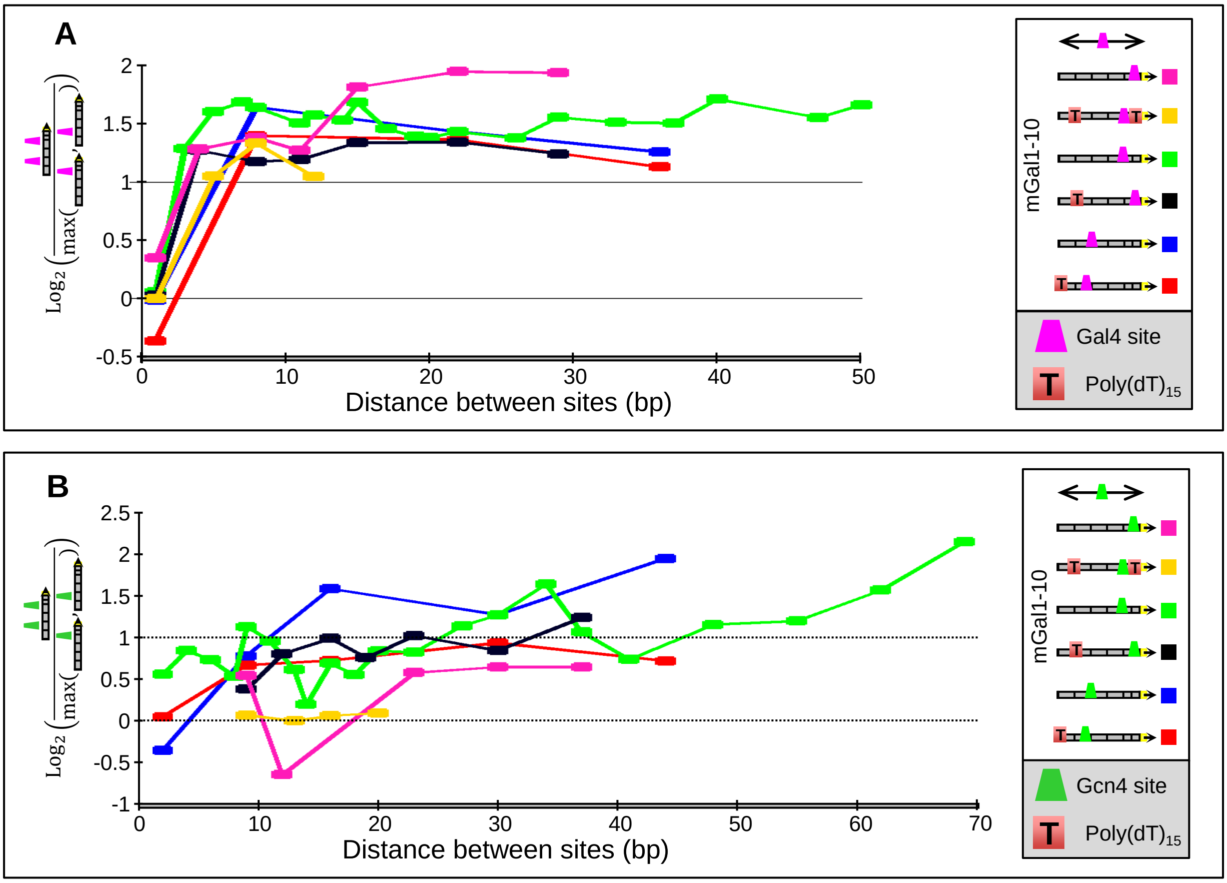
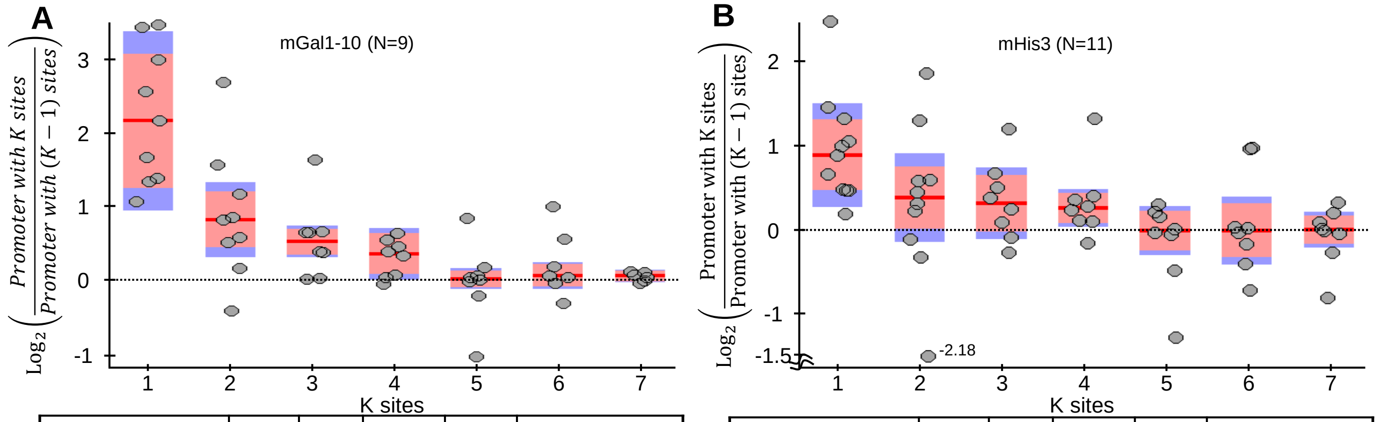

A
Log2(max(,))
2
1.5
1
0.5
0
-0.5
0
10
20
30
40
50
Distance between sites (bp)
mGal1-10
T
T
T
T
Gal4 site
T
Poly(dT)15
B
Log2(max(,))
2.5
2
1.5
1
0.5
0
-0.5
-1
0
10
20
30
40
50
60
70
Distance between sites (bp)
mGal1-10
T
T
T
T
Gcn4 site
T
Poly(dT)15

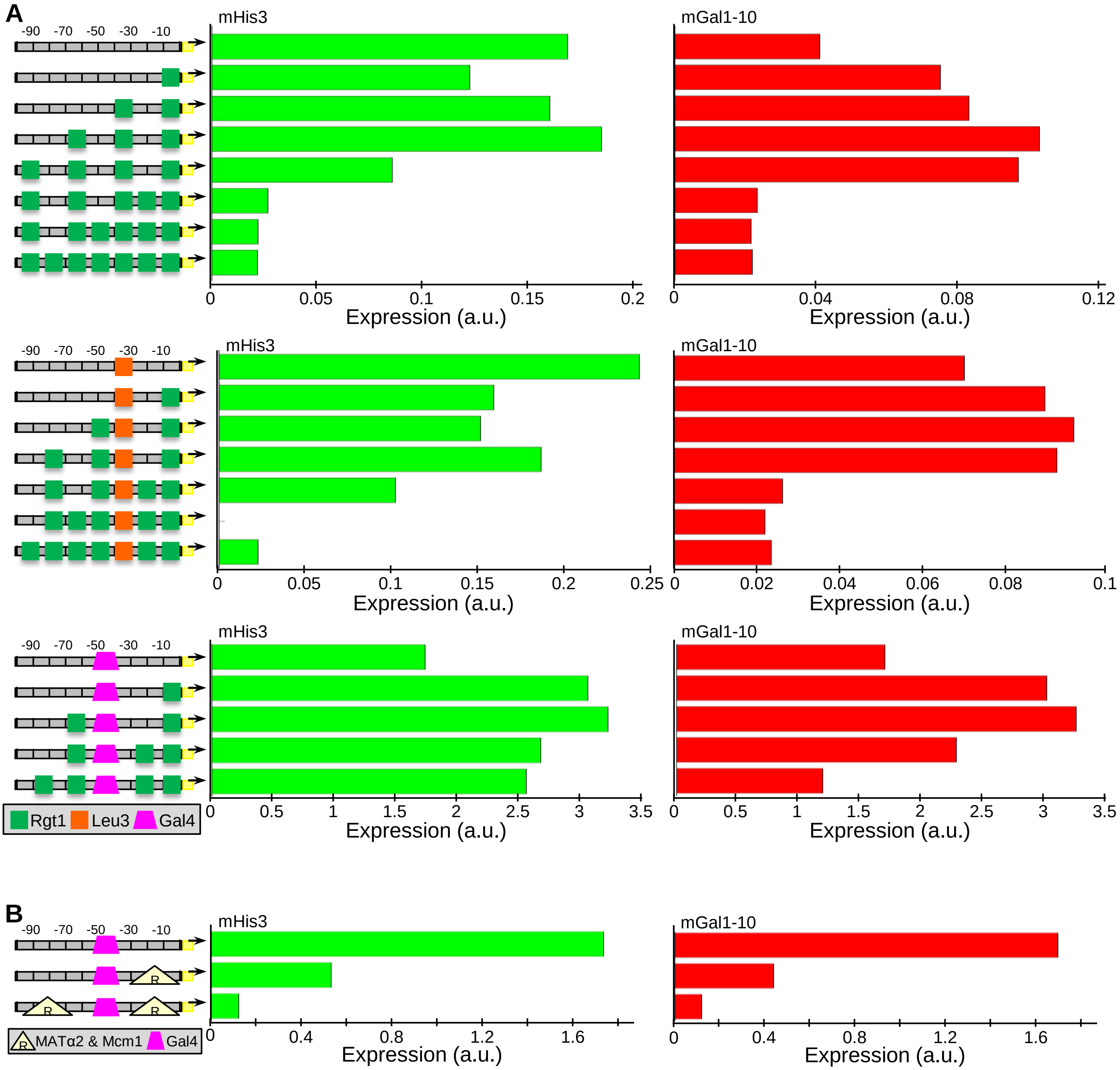
Figure S20



Transcription factors	Rap1	Bas1	Leu3	Swi4	Fhl1
Gcn4 (weak affinity)	Gcn4	Rsc3	Met31/2	Gcr1	Bas1 weak affinity

Transcription factors	Rap1	Bas1	Leu3	Swi4	Fhl1
Gcn4 (weak affinity)	Gcn4	Rsc3	Met31/2	Gcr1	Bas1 weak affinity

Figure S21


A
-90 -70 -50 -30 -10
mHis3
0
0.05
0.1
0.15
0.2
Expression (a.u.)
mGal1-10
0
0.04
0.08
0.12
Expression (a.u.)
-90 -70 -50 -30 -10
mHis3
0
0.05
0.1
0.15
0.2
0.25
Expression (a.u.)
mGal1-10
0
0.02
0.04
0.06
0.08
0.1
Expression (a.u.)
-90 -70 -50 -30 -10
mHis3
0
0.5
1
1.5
2
2.5
3
3.5
Expression (a.u.)
mGal1-10
0
0.5
1
1.5
2
2.5
3
3.5
Expression (a.u.)
🟩Rgt1 🟧Leu3 Gal4
B
-90 -70 -50 -30 -10
R
R
R
mHis3
0
0.4
0.8
1.2
1.6
Expression (a.u.)
mGal1-10
0
0.4
0.8
1.2
1.6
Expression (a.u.)
R MATα2 & Mcm1 Gal4