Unraveling determinants of transcription factor binding outside the core binding site

Michal Levo*[1,2], Einat Zalckvar*[1,2], Eilon Sharon[1], Ana Carolina Dantas Machado[3], Yael Kalma[2], Maya Lotam-Pompan[2], Adina Weinberger[1,2], Zohar Yakhini[4,5], Remo Rohs[3] and Eran Segal[1,2]

* These authors contributed equally to this work, and are listed alphabetically.

[1] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel

[2] Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel

[3] Molecular and Computational Biology Program, Departments of Biological Sciences, Chemistry, Physics, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA

[4] Computer Science Department, Technion – Israel Institute of Technology, Haifa 32000, Israel.

[5] Agilent Laboratories, Santa Clara, CA 95051, USA

Correspondence to: Eran Segal[1,2] email: eran.segal@weizmann.ac.il

Supplementary information

A. Experimental Procedures

A1. Proteins

Gcn4

GST-His-GCN4 (1-109, *S. cerevisiae*) was cloned into pET-TevH plasmid and expressed in Bl21(DE3) bacteria at 15°C over night following induction with 0.2 mM IPTG. The cells were suspended in PBS supplemented with 20% Glycerol, 5 mM EDTA, 2 mM DTT, protease inhibitor cocktail (Mercury) and PMSF and lysed by sconication. The cell debri was removed by centrifugation and the soup was incubated with 3ml Glutathione Sepharose beads (GE Helthcare) for 1 h. The beads were washed with PBS containing 20% Glycerol and 5mM EDTA and the bound protein was eluted from the beads with 2 ml buffer containing 100mM Tris pH 8, 20% Glycerol, 5 mM EDTA, 2 mM DTT and 20 mM reduced Glutathion. The pooled fractions containing GST-GCN4 were applied to an ion exchange column (Tricorn Q 10/100 GL, GE Healthcare) equilibrated with 50mM Tris pH 7.5 and the protein was eluted with a linear gradient containing 1 M NaCl. TEV protease was added to the pooled fractions and left over night at 4°C to cleave the GST tag. The soup was then incubated consecutively for 1hr with Glutathione Sepharose beads (to remove the GST tag) and then with Ni NTA beads (to remove the His-TEV protease). 20% glycerol was added to the unbound protein containing GCN4. GCN4 was finally filtered through a centricon with a 30 kDa MWCO (Millipore) and re-applied to a second centricon with a 3kDa MWCO (Millipore). In this step, the protein was concentrated to 0.4 mg/ml and flash frozen with liquid nitrogen.

Gal4

Gal4 (1-147, *S.cerevisiae*) + α helix, 0.5 mg/ml (abcam).

A2. Protein/DNA molar ratios used in the application of BunDLE-seq

TF concentrations chosen range from the minimal concentration yielding sufficient binding for isolation of bound DNA, up to a concentration with a minimal non-specific binding (as assayed by comparing the binding of a sequence containing the tested TF binding site to a sequence lacking this site).

	Core TFBS library	Fixed-flanks TFBS library
Gcn4/DNA	No protein, 1:8, 1:7, 1:6, 1:5, 1:4, 1:3, 1:2, 1:1	No protein, 1:5, 1:3, 3:1
Gal4/DNA	No protein, 1:1, 3:1, 5:1, 10:1	No protein, 1:1, 3:1

Histone octamers/DNA - 3:1 protein/DNA molar ratio.

A3. Library description and preparation

A library of 6,500 sequences of 150 bp in length, as described elsewhere(Sharon et al. 2012), was used as input for binding measurements. Among these sequences ~3800 contained at least one binding site for either Gcn4 or Gal4 or served as controls. An additional library of 13,000 sequences of length 200bp was also used. Among these sequences ~7700 contained sites for Gcn4 or Gal4 with fixed flanks.

Each library was synthesized by Agilent(LeProust et al. 2010) and cloned into the pKT103-based plasmid as described elsewhere(Sharon et al. 2012). Input sequences for BunDLE-seq were produced by 11-12 PCR amplification cycles on the pKT103-based plasmid using Herculase II Fusion DNA Polymerase (Agilent) in 96 wells plates. The primers used amplify the pooled DNA were 5'-GGGGACCAGGTGCCGTAA-3' (forward primer) and 5'-TTATGTGATAATGCCTAGGATCGC-3' (reverse primer). The PCR products from the 96 wells plates were joined and concentrated using Amicon Ultra, 0.5 ml 30 K tubes (Merck Millipore). The concentrated DNA was then run on 2.5% agarose gel with crystal violet

(Sigma), and the DNA was cut from the gel and purified using QIAquick gel purification kit (QIAGEN) while using the supplied QG buffer at room temperature (to minimize GC bias(Quail et al. 2009)).

Commonly used TFBS:

- Strong Gcn4 9 bp site: ATGACTCAT
- Strong Gcn4 23 bp site: TCTCCATATGACTCATAAATATA
- Strong Gcn4 29 bp site: TCCTCTCCATATGACTCATAAATATAGTA
- Weak Gcn4 9 bp site: ATGACTCGT
- Strong Gal4 17 bp site: CGGAAGACTCTCCTCCG
- Strong Gal4 23 bp site: AGACGGAAGACTCTCCTCCGTGA
- Strong Gal4 29 bp site: CTGAGACGGAAGACTCTCCTCCGTGAGTC
- Weak Gal4 17 bp site: CGCGCCGCACTGCTCCG

Commonly used sequences contexts:

- *GAL1_10*_NULL:

AAGCCGGACAGGCAGCGACAGCCCTGACAGACAAGACTCTCCTAGCTGCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTACAGCCGCAC

- *HIS3*_NULL:

GCACTAAATCGGAACCCTAAAGGGAGCCCCCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGGCGAGAAAGGAAGGGAAGAAAGCGCCACCTAGCGGAA

A4. Brief description of the expression measurements

Fully designed DNA libraries at length of 150bp or 200bp were synthesized by Agilent. The DNA was cloned into a low-copy pKT103-based plasmid upstream of a constant yellow fluorescent (YFP) reporter gene located adjacent to a mCherry-encoding gene under a constitutive TEF2 promoter. The plasmids were then amplified in Escherichia coli and transformed into yeast cells.

Yeast library cells were sorted by FACSAria cell sorter (Becton-Dickinson) at mid-exponential phase (OD600 0.5–1.5) according to the ratio of YFP and mCherry, thereby normalizing for extrinsic noise effects (mCherry expression was used for gating to enrich a population with a single plasmid). The cells were sorted three times recursively into four bins, producing a total of 64 bins. An alternative sorting scheme, in which cells were sorted directly into 16 bins, yielded highly similar results (R^2=0.95). Cells from each bin were grown to stationary phase and one million cells from each bin were taken for colony PCR using primers corresponding to the promoter region of the plasmid; with the 5′ primer containing a unique 5-bp barcode sequence that was specific to each bin. All PCR products were joined and sent to sequencing by the SOLiD system.

Based on a mean expression value obtained for each expression bin and the fraction of cells with each promoter found in each expression bin (obtained from the sequences reads) a mean expression value per promoter sequence was computed; Namely, 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.

More details on the expression measurements can be found in Sharon and Kalma et al.(Sharon et al. 2012)

A5. Unique features and limitations of BunDLE-seq

This work aims to contribute to a recently emerging view of TF binding that goes beyond the sole characterization of core TF sites. While several recent studied, discussed in the main text, provide progress in this direction, we are still far from bridging the current gap between ‘classical’ in-vitro characterizations of TF binding to a quantitative account of TF binding events in-vivo, with the latter being paramount for a quantitative understanding of gene expression. This challenge can be addressed by the development of novel methodologies, overcoming properties of commonly used *in-vitro* assays that render them distinct from the in-vivo environment. In that respect our approach possesses an important unique feature, that is, the usage of long (up to 200 bp) DNA templates. This allows us to determine, in TF-specific manner, what is the number of bps beyond the core motif that seem to bear information that can influence TF binding (for both TF tested, we find this information extends beyond previous characterizations, see discussion in the text). The length of the examined templates, coupled with two additional unique properties of our assay, namely the designed nature of our sequences, as well as the ability to isolate (via gel shift) the occurrence of single versus co-occurring binding events, allows us to extend the investigation from that of single site containing sequences to that of multiple sites containing sequences. By doing so we are able to highlight the contribution of features that are commonly disregarded when in-vitro deduced binding preferences are utilized in the interpretation of regulator sequences; For instance, the simple count of motif occurrences that is often performed disregards the differential likelihood of co-occurring binding events, stemming from differential location of the site relative to one another, that is captured by our assay.

The length of our examined sequences further facilities a complementary experiment, carried out with TF's prominent competitors for DNA binding within cells, namely histones (naturally, other approaches employing templates significantly shorter than 147 bp are inherently incapable of examining this aspect). Here too, the designed, rather than the commonly used random nature of our templates, provides means to explicitly test the effect of prevalent sequences in eukaryotic promoters (e.g., poly(dA:dT) tracts in varying length) as well as the effect of different native promoter contexts.

These unique features of our approach provide means to widen the scope of TF binding investigation from a local, site-oriented perspective to a regulatory-sequence based perspective. The commonly performed, extensive surveys of *k*-mers are replaced here by the systematic examination of DNA templates that more closely resemble, both in terms of length and composition, regulatory sequences. This enable the usage of same set of sequences as promoters in a reporter activity assay; thereby facilitating a 'controlled' comparison of in-vitro binding measurements versus *in-vivo* expression measurements (see discussion in the main text and supplementary section D).

It should be noted, however, that despite the advantages mentioned above our approach has several limitation. These include limitation stemming from the technology employed to synthesize the examined sequences(LeProust et al. 2010); most notably, the current limitation on sequences length (overall ~200 bp, with the variable region limited to ~160 bp, due to constant edges used for amplification). This naturally cannot encompass the entire complexity of native eukaryotic regulatory sequences. Furthermore, the number of synthesized sequences is also limited (in this study >10,000 were examined) thereby limiting the number of parameters that be examined systematically; For instance, here, we have dedicated many variants for the examination of a TFBSs' location and multiplicity, but this was generally done only on two sequence contexts. Notably, extensive sequence mutagenesis (single, double, triple permutations and so on), beneficial in the characterization of binding affinities, can rapidly saturate a seemingly large pool of sequences.

Future applications of our approach can aim to improve and extend additional aspects; for instance, as in SELEX-based assays, performing several rounds of selection can increase the dynamic range for which the method provides accurate binding quantifications.

Additionally, while current applications of BunDLE-seq were carried out with a single type of binding protein (either TFs or histones), future applications can be done in the presence of multiple types of proteins. The ability to separate different binding events on the gel would be particularly beneficial in such applications.

B. Measures employed

The sequencing data provides, for each tested sequence, its frequency in each of the bands. An additional sample, treated as all the other just with no exposure to the TF (DNA only), served to estimate the frequency of each sequence in the initial pool.

The ‘binding score’ computed per sequence, per band, is the frequency of that sequence among the sequences extracted from that band divided by its frequency in the initial library.

Following the annotation from(Zhao et al. 2009), the probability of binding of sequence i ($p(s = 1|S_i)$) can be computed, using Bayes’ rule, by the proportion of this sequence in the bound band ($p(S_i|s = 1)$) multiplied by the probability of that band ($p(s = 1) = \sum_j p(s = 1|S_j)P(S_j)$), divided by the proportion of the sequence in the initial library pool ($p(S_i)$).

$$p(s = 1|S_i) = \frac{p(S_i|s = 1)p(s = 1)}{p(S_i)}$$

From the DNA extracted from the TF bound band we had $p(S_i|s = 1)$ and from the band with the initial library we had $p(S_i)$. Our computed binding score, per sequence i in band ‘1’ is $\frac{p(S_i|s=1)}{p(S_i)}$.

It is thus proportional to the probability of binding of sequence i $p(s = 1|S_i)$.

$$binding\_score(sequence\ i, band\ 1) = \frac{p(S_i|s = 1)}{p(S_i)} = \frac{p(s = 1|S_i)}{p(s = 1)}$$

Thus, any analysis that compares the binding score of different sequences within the same band essentially compares their probability of binding multiplied by ‘constant’ (that is the probability of that band).

Specifically, computing the ratio of binding scores for two sequences is equal to computing the ratio of their probability to be bound (as the probability of the bound band cancels out).

$$\frac{binding\_score(sequence\ i, band\ 1)}{binding\_score(sequence\ j, band\ 1)} = \frac{p(S_i|s = 1)}{p(S_i)} \Big/ \frac{p(S_j|s = 1)}{p(S_j)} = \frac{p(s = 1|S_i)}{p(s = 1)} \Big/ \frac{p(s = 1|S_j)}{p(s = 1)} = \frac{p(s = 1|S_i)}{p(s = 1|S_j)}$$

Notably, the probability of a sequence to be bound by its regulators can be used, as was done in previous studies, as a proxy to the transcription outcome (e.g., under a simplifying assumption that in some regimes this probability is proportional to the binding probability of RNA polymerase, which is, in turn, proportional to the resulting transcription rate)(Raveh-Sadka et al. 2009; Raveh-Sadka et al. 2012; Sharon et al. 2012). Thus, computing the probability of a sequence to be bound by its regulators, or a measure proportional to it as the 'binding score', facilitates the comparison of our binding measurements to corresponding expression measurements.

Additionally, we find this type of measure is very intuitive and easily interpretable when examining sequences with multiple binding sites and particularly when considering co-occurring binding events (for instance compared to affinity-type of measurement, e.g., the probability of a sequence with two binding sites to be bound simultaneously at both sites is more intuitive than the notion of the affinity of this sequence to two TF molecules). Related to this, is the fact that this type of measure is easily amendable for thermodynamic modeling of binding (see section below), thus allowing us to examine our quantitative and predicative understanding of the data. Finally, we find that that the 'binding score' measure is extremely robust in our system (see for instance its high reproducibility in Figure S1). We therefore choose to use this measure in the analyses throughout this work, unless specified otherwise (including the analyses pertaining to nucleosome formation).

C. Detailed description of the thermodynamic model:

General description:

Under the assumption of a thermodynamic equilibrium we can compute, based on the Boltzmann distribution the probability of each binding state for any given sequences (i.e. the probability that the sequence is not bound, bound by one TF molecule, bound by two TF molecules or a higher number of TFs.). We do so by summing the statistical weight associated with configurations that contribute to such a state and dividing by the sum of statistical weights of the configurations contributing to any possible state. This can be written:

$$P(state) = \frac{\sum_{c \in State} W(c)}{\sum_{c' \in C} W(c')}$$

where $\boldsymbol{W(c)}$ represents the statistical weight associated with configuration *c* (and *C* represents the group of all possible configurations).

The definition of W(c):

Let us consider a configuration in which *k* TF molecules are bound to specific binding sites, and denote $\boldsymbol{w_i}$ as the weight contribution from a TF binding event. $\boldsymbol{w_i}$ is modeled as the multiplication of the TF concentration ($\boldsymbol{\tau_{TF}}$) and the affinity or energetic gain from the binding of a single molecule to site *i* ($\boldsymbol{F_{(TF,site_i)}}$).(Raveh-Sadka et al. 2009)

This can thus be written as:

$$W(c) = F_0 \prod_{i=1}^{k} \tau_{TF} \cdot F_{(TF,site_i)} = F_0 \prod_{i=1}^{k} w_i$$

Let us employ the simplifying assumption that the affinity of the TF to all sites is the same. We can thus denote any $\boldsymbol{w_i}$ as $\boldsymbol{w}$.

$$W(c) = F_0 \prod_{i=1}^{k} w_i = F_0 \cdot w^k$$

Let us consider a sequence with s binding sites; there are $\binom{\boldsymbol{s}}{\boldsymbol{k}}$ configurations in which k sites among the s sites are bound.

Thus:

$$p(k\_\,TF\_bound\_in\_a\_sequenc\_with\_s\_sites) = \frac{\sum_{c \in k\_sites\_bound} W(c)}{\sum_{c' \in C} W(c')} = \frac{\binom{s}{k} w^k}{1 + \sum_{i=1}^{s} \binom{s}{i} w^i}$$

Notably, since F_0 appears in all configurations, it cancels out in this computation.

Application to our sequences:

A set of sequences with all combinations of one to seven binding sites for Gcn4, in seven predefined locations, were considered. We assumed that these sequences could be found in each of eight types of states, either not bound by the TF, or bound by one to seven TF molecules.

We performed eight experiments with different concentration of the Gcn4, and extracted the DNA from each of the formed bands on the gel (both the bands representing unbound DNA, and the bands representing bound DNA).

We computed, for each of the sequences in the examined set, its ‘binding score’ in each of the bands (that is its frequency among the DNA extracted from that band divided by its expected frequency in the initial pool). We then computed an average score across the sequences sharing the same number of sites (we further normalized these curves by dividing them by the mean of the averaged score computed to sequences with four sites and sequences computed with five sites). These curves are plotted in blue in Figure 5B.

We then employed the thermodynamic model described above to compute a corresponding measure, that is $p(k\_\,TF\_bound\_in\_a\_sequenc\_with\_s\_sites)$ (normalized by the mean values for sequences with four and five sites). For compression to the experimental curve produced for the naked DNA band k = 0, for the one TF bound band k = 1 and for the additional band formed in high TF concentrations, likely representing DNA bound by two TF molecules, we used k = 2. Notably, as we are producing predictions for the average score for a sequence with s sites (averaging over the exact site's locations), the simplifying assumption employed in the model, that the affinity of the TF to all sites is the same, is a reasonable one.

We computed the model predictions for a range of possible values for the parameter w (from 0.0001 to 2 in jumps of 0.0001). For each of the eight experiments, and for each assignment for the w parameter, we computed the sum of the L2 distances between the experimental curves computed for the bands formed on the gel in this experiment and the corresponding curves predicted by the model. We then chose, per experiment, the value for w that yielded the curves with the lowest sum of L2 distance.

The best fitted curve was plotted in black in figure 5B, and the w values with which it was computed are plotted on the y axis of figure 5C, against the concentrations of Gcn4 used in the eight experiments.

D. Extended discussion on binding-to-expression comparisons:

The binding measurements of Gcn4, Gal4 and histones were carried out also on 6500 designed sequences that recently served as promoters, placed upstream of a fluorescence reporter, in a high-throughput reporter activity assay (herein after referred to as 'expression measurements')(Sharon et al. 2012). It is thus of interest to examine the outcomes obtained as the level of binding with respect to those obtained at the level of expression.

When we examine our Gcn4 binding measurements we focus our attention on the 1050 sequences, among the 6500 for which expression measurements were carried out, that were designed to contain only Gcn4 sites. Naturally, the expression outcome of other sequences, designed with additional binding sites for other TFs, is expected to reflect the activity of the multiple regulators involved, and is therefore less comparable to binding measurements carried out with only Gcn4.

We found the Pearson's correlation between the expression measurements for these 1050 sequences, and our Gcn4 binding score (computed based on a band representing a single Gcn4 binding event, in an experiment carried out with an intermediate concentrations of Gcn4, that is the one used in all the single concentration-based analyses presented), is 0.638 (Figure S12A). However, despite the fact that these sequences were designed not to include additional TF sites, they can definitely serve, within cells, as templates for the formation of nucleosomes. Our binding measurements, carried out with histones rather than the regulating TF, demonstrate that the propensity of these sequences to form nucleosomes can vary. These measurements capture the previously discussed nucleosome disfavoring nature of poly(dA:dT) tracts (Figure 6B,C)(Struhl and Segal 2013); suggesting for instance, that two sequences differing in the presence of a poly(dA:dT) tract (that is not located immediately adjacent to the TF site) could have a similar Gcn4 binding score in our in-vitro measurements would show differential Gcn4 binding within the cell, and therefore differential expression, since they are expected to differ in their nucleosome occupancy (consistent with previous studies(Raveh-Sadka et al. 2012)). Indeed, when we remove sequences with such tracts from our analyses, the Pearson's correlation between expression and Gcn4 binding score increases to 0.711 (more on the role of these tracts with respect to both TF binding and nucleosome binding can be found in the main text and figures 4H,6B-C,S11). Similarly, our measurements reveal a differential propensity to form nucleosomes between sequences that were designed with the *HIS3*-derivedcontext and sequences with the *GAL1-10*-derived context (with the latter showing higher nucleosome occupancy, consistent with their measured lower expression, see figure 6D,E). While the differential expression between these two contexts can stem form a combination of multiple mechanisms, the possible contribution of nucleosomes emerging form our measurement, suggests Gcn4 binding would be more predictive of expression on each of these contexts separately rather than when mixed together; indeed we find the Pearson's correlations on these two main contexts increase to 0.868 and 0.883 (for the *HIS3*-derived and the *GAL1-10*-derived sequences respectively, Figure S12B). It should be noted that the magnitude of the increase in the Pearson's correlation upon the sub-classifications of these sequences based on different sequence features should not be regarded as a quantitative measure of these features contribution to expression, for instance because these features were not originally designed to be similarly represented among the tested sequences.

If we further examine the obtained Pearson's correlation, we see that as expected it is highly influenced by the number of Gcn4 binding sites (supplementary figure S12,C,D); in both expression and binding measurements an increase in the number of Gcn4 binding sites results in an increase of the measured value. However it should be noted this Pearson's correlation is somewhat misleading since the events accounted for by these measurements are different. Specifically, in our binding measurements we chose to isolate different types of binding configurations, namely single binding events and two co-occurring binding events (distinguishing between these binding events allowed us to perform the analysis presented in figure 5). The Gcn4 binding score we utilize here, in these comparisons to expression measurements, refers to the probability for a single binding event, and thus only accounts for type of configuration that can take place within the cell on a sequence containing multiple binding sites.

We thus further zoom-in on sequences with a single Gcn4 binding sites, and find that for these sequences the Pearson's correlations between binding and expression is 0.819 and 0.75, on the *HIS3*-derived and the *GAL1-10*-derived sequences respectively (Figure S12E,F). Notably, these sequences include sequences with a weak core site ('ATGACTCGT') or sequences with a strong core site ('ATGACTCAT', except for 1-2 sequences on each of the promoter contexts with the reverse complement site 'ATGAGTCAT'). These difference in site affinity are captured by our binding methods and they also seem to play a role in the expression outcome, thus contributing to the observe Pearson's correlation (Figure S12E,F). The relation between our *in vitro* binding and expression obtained when we further zoom in on the sequences with a single strong core site ('ATGACTCAT') differentially located along the sequence can is shown in figures S4A and B (that are also discussed in the main text); This correlation (0.62 and 0.5 for the *HIS3* and *GAL1-10*-derived contexts respectively) is particularly interesting as it suggests that the interaction of the TF and DNA can already contribute to differential expression upon differential site location (possibly through the sequences flanking the core site), even before accounting for more commonly suggested mechanisms, involving for instance the interaction between the TF and the transcription machinery.

We performed a similar comparison between our Gal4 binding measurements and the expression measurements of the relevant sequences (Figure S12G-L). For this TF we observe a general preference for binding to sequences with the *HIS3*-derived context compared to sequences based on the *GAL1-10*-derived context, that can contribute to the generally higher expression obtained with the first context compared to the latter, in addition to the contribution of differential nucleosomes formation on these contexts.

As in the case of Gcn4, here too, we are able to distinguish in our binding measurements between weak core sites (‘CGGAGCAGTGCGGCGCG’) to strong sites (with ‘CGGAAGACTCTCCTCCG’ commonly used, except for 1-2 sequences on each promoter context with the reverse complement site ‘CGGAGGAGAGTCTTCCG). When we zoom in on sequences with the same strong core site, differentially located, we obtain the Pearson’s correlations presented in figures S4C and D (for the *HIS3* and *GAL1-10*-derived sequences respectively) demonstrating, that for this TF as well, sequence determinants outside the core site can contribute to differential expression.

Supplementary Figures

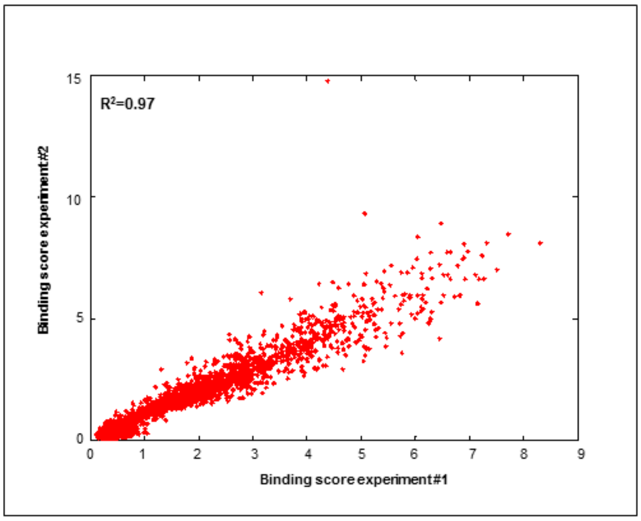


Figure S1. High reproducibility between replicates. Scatter plot of binding scores obtained in two experimental replicates performed using Gcn4 and 6,500 sequences of length 150 bp. Binding, isolation and amplification of the DNA extracted from each band, and sequencing were performed independently.

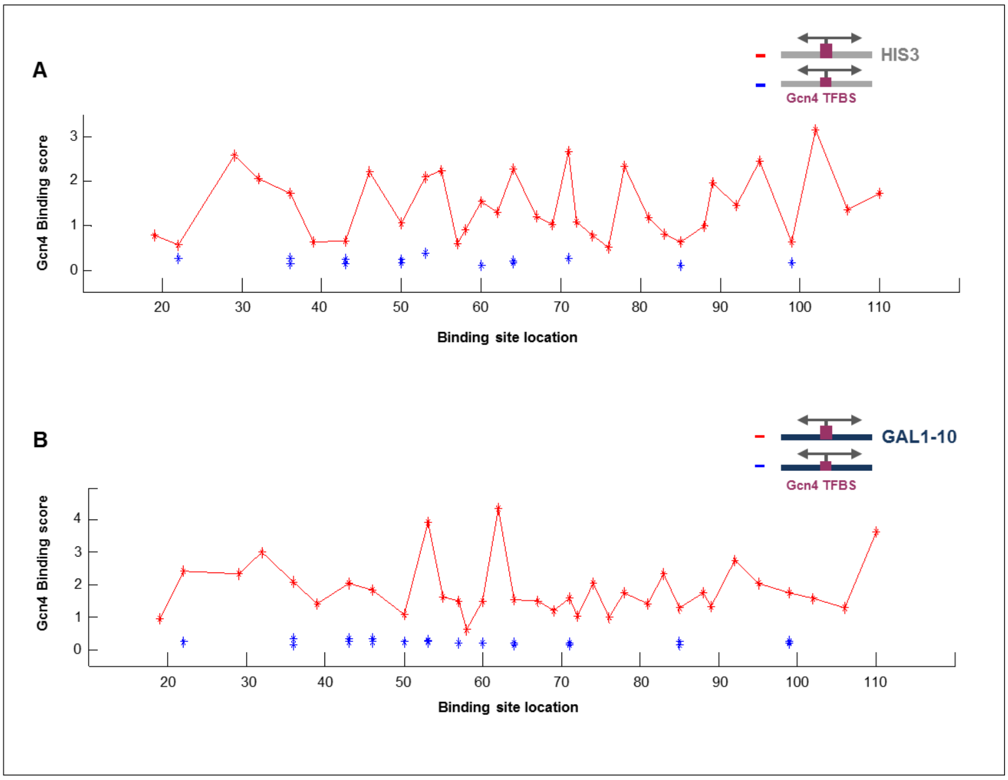


Figure S2. The effect of the TFBS location on TF binding spans a comparable range to that attained by reducing binding site affinity by mutation to the TFBS. For a set of sequences in which a single strong (in red) or weak (in blue) binding site was placed at different locations along a specific context, plotted is log2 of the ratio of the binding score attained by each sequence (with the x coordinate marking the location of the center of the site) divided by the median binding score across all sequences in this set. (A) A Gcn4 TFBS of 9 bp placed along the *HIS3*-derived context. (B) A Gcn4 TFBS of 9 bp placed along the *GAL1-10*-derived context.

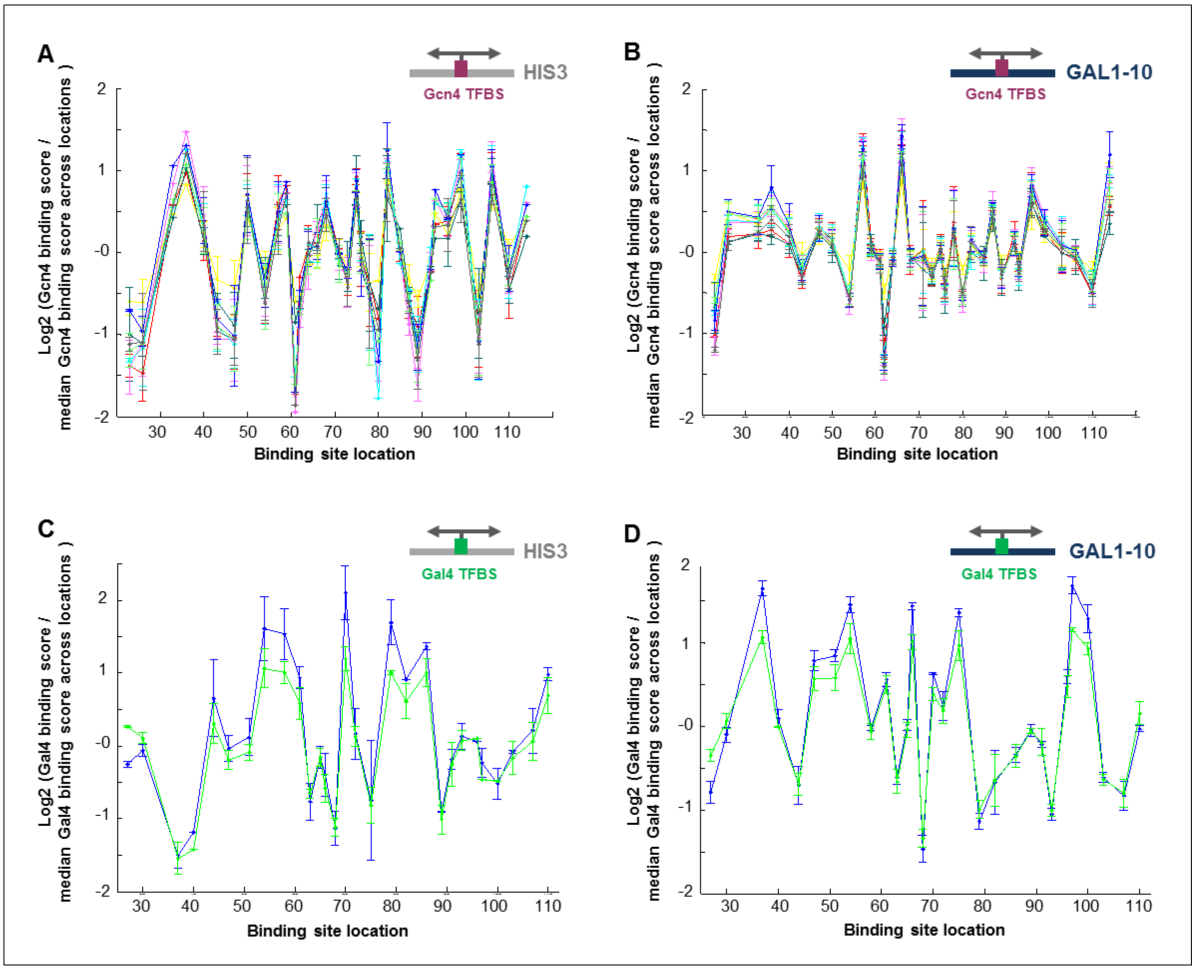


Figure S3. TFBS location effects are conserved across different TF concentrations. For a set of sequences in which a single strong binding site was placed at different locations along a specific context, plotted is log2 of the ratio of the binding score attained by each sequence

(with the x coordinate marking the location of the center of the site) divided by the median binding score across all sequences in this set, and across different concentrations (different graph colors). (A) A Gcn4 TFBS of 9 bp placed along the *HIS3*-derived context. (B) A Gcn4 TFBS of 9 bp placed along the *GAL1-10*-derived context. (C) A Gal4 TFBS of 17 bp placed along the *HIS3*-derived context. (D) A Gal4 TFBS of 17 bp placed along the *GAL1-10*-derived context.

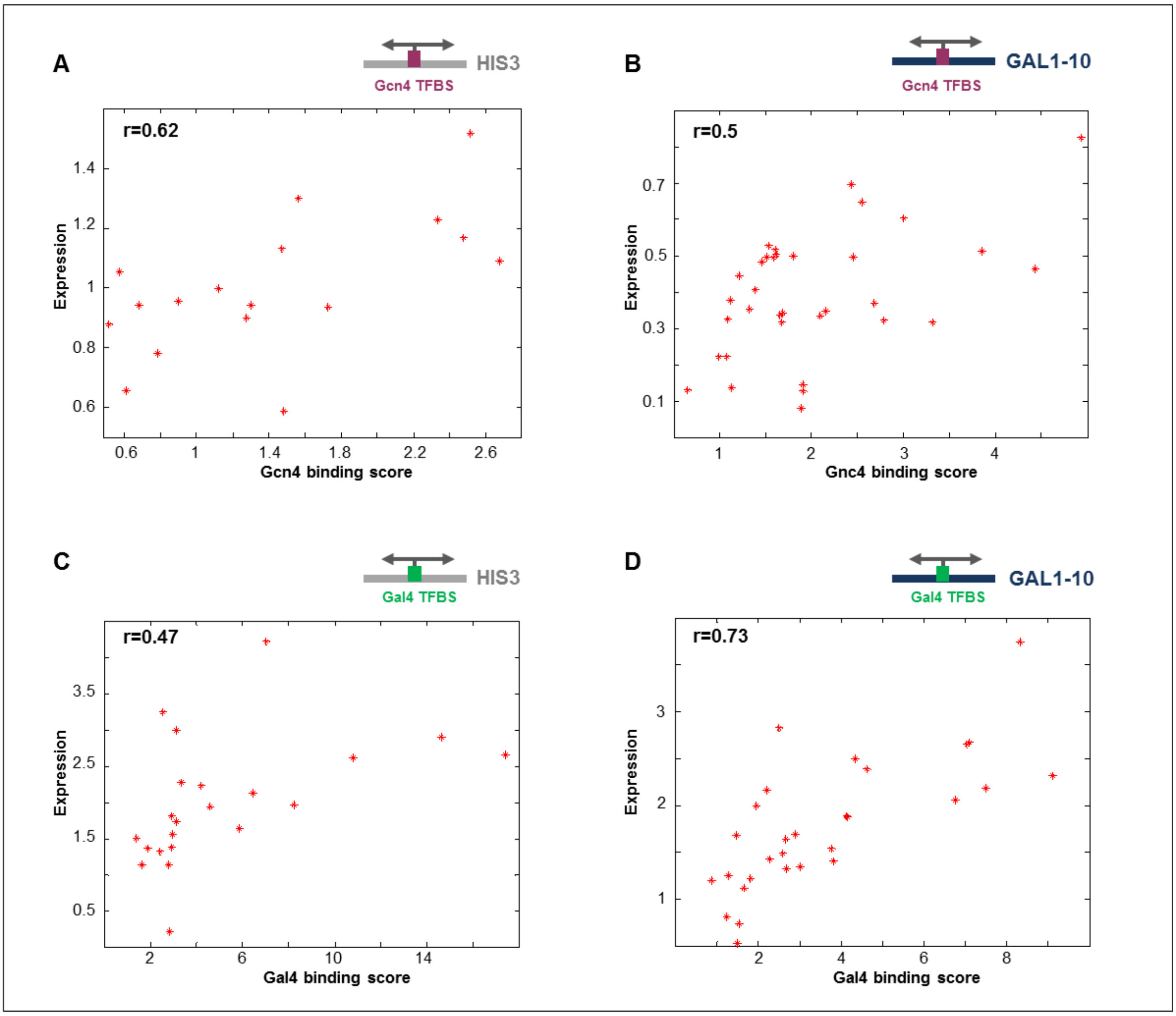


Figure S4. The effect of site location on TF binding might contribute to differential expression. Scatter plot of binding score versus expression measurements(Sharon et al. 2012) of sequences in which a single strong binding site was placed at different locations along a

specific sequence context. (A) A Gcn4 TFBS of 9 bp placed along the *HIS3*-derived context (17 seq, with no other regulatory element). (B) A Gcn4 TFBS of 9 bp placed along the *GAL1-10*-derived context (34 seq, with no other regulatory element). (C) A Gal4 TFBS of 17 bp placed along the *HIS3*-derived context (23 seq, with no other regulatory element). (D) A Gal4 TFBS of 17 bp placed along the *GAL1-10*-derived context (29 seq, with no other regulatory element). (r = Pearson's correlation).

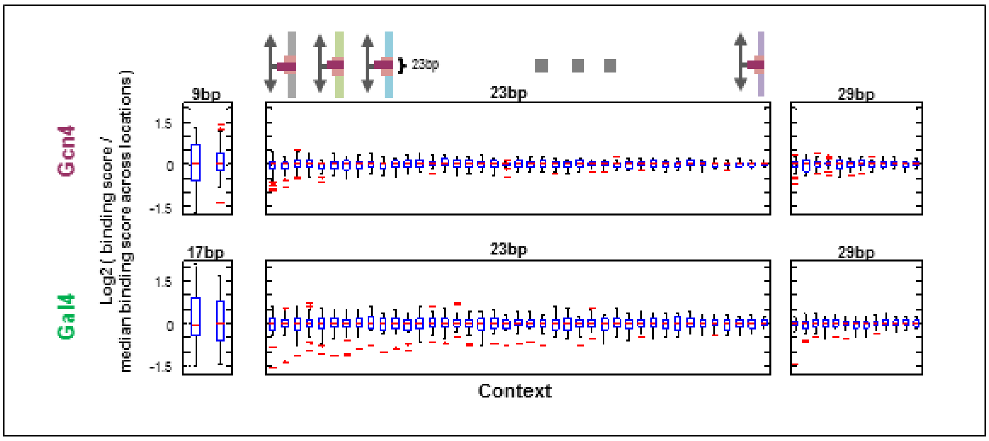


Figure S5. Differential placement of the core TFBS results in more pronounced fluctuations in binding than differential placement of a fixed flanks TFBS. For more than 40 different contexts (including the *HIS3*-derived and *GAL1-10*-derived context and additional random contexts), shown is a boxplot of the log2 ratio of binding scores at different site locations divided by the median score across all locations; upper panels-Gcn4 (9 bp core site, or 23 bp/29 bp fixed-flanks site), lower panels-Gal4 (17 bp core site, or 23 bp/29 bp fixed-flanks site). Across all of these contexts the magnitude of binding fluctuations when moving a TFBS with fixed flanks was always significantly lower than the magnitude of fluctuations attained when only the core binding site was moved (along the *HIS3*-derived and *GAL1-10*-derived contexts), thus also changing its flanking base pairs.

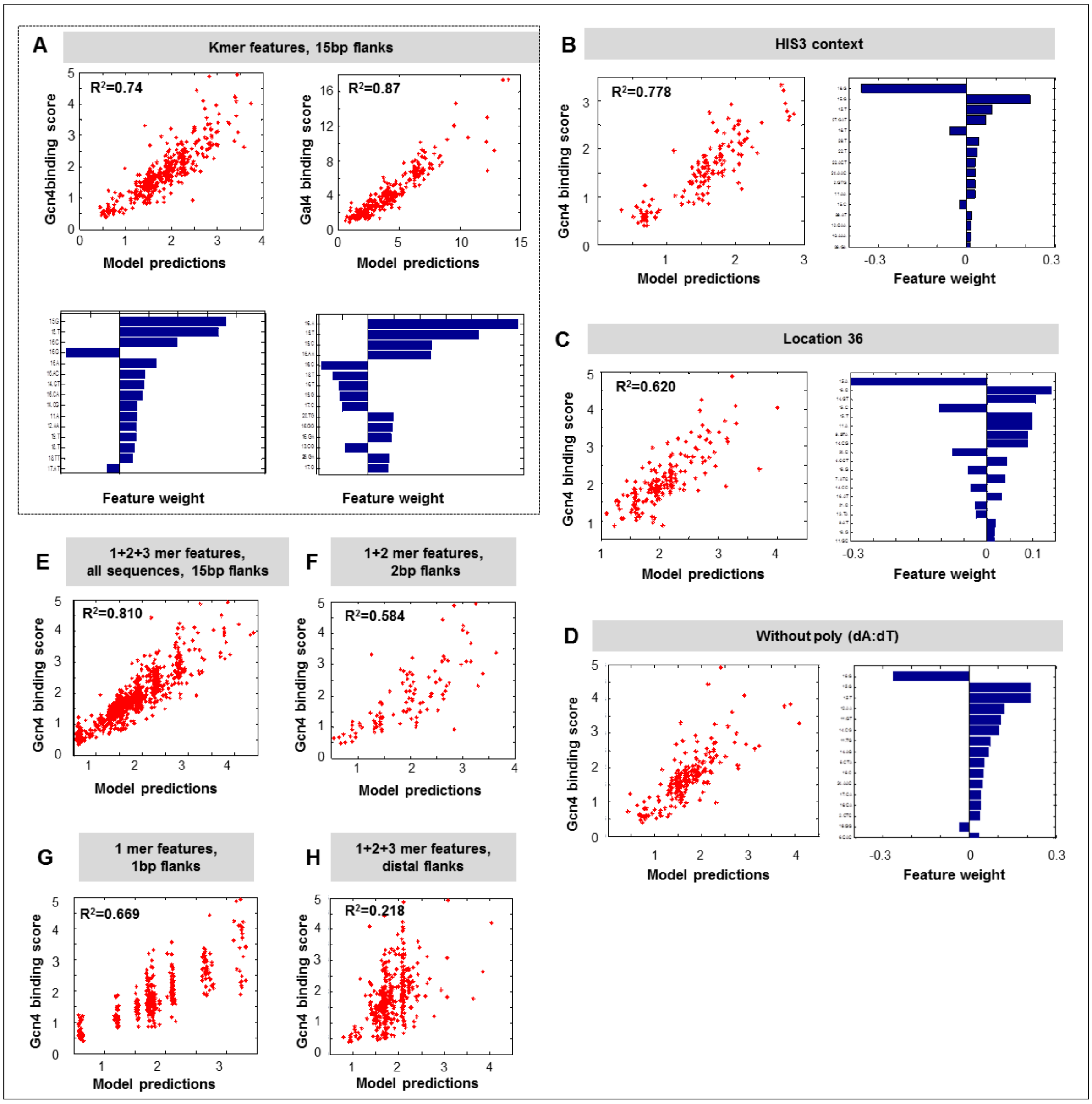
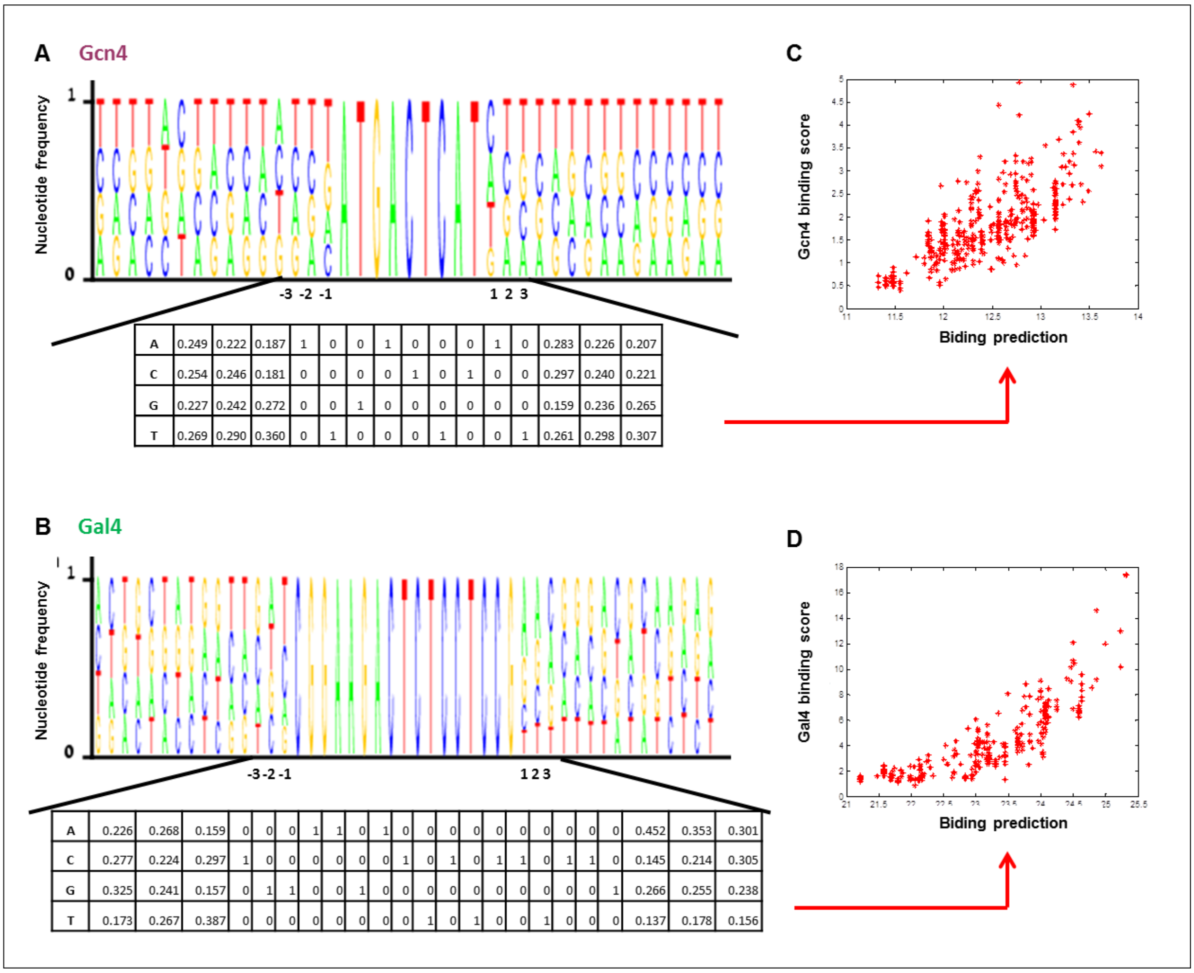


Figure S6. Computational models based on the nucleotide content of the flanking sequences predict differential binding of sequences containing the same core TFBS. Scatter plots of Gnc4 binding score versus model predictions, with a 15 bp flanks, 1-mers+2mers k-mer count model on 412 sequences (left), and Gal binding score versus model predictions with a 15 bp flanks, 1-mers+2mers k-mer count model on 315 sequences (right), with the weights of the top 15 sequence features below the corresponding graph. (B,C,D) To further verify that the design of our sequences does not introduce significant confounding effects, we run the model also on several subsets of the sequences controlling for potential biases: (B) 15 bp flanks, 1-

mers+2-mers+3-mers on ~160 sequences containing the *HIS3* context. (C) 15 bp flanks, 1-mers+2-mers+3-mers, on ~190 sequences containing Gcn4 strong site at location 36. (D) 15 bp flanks, 1-mers+2-mers+3-mers on 217 sequences excluding poly(dA:dT)-containing sequences. (E,F,G,H) Scatter plots of Gnc4 binding score versus model predictions, with the following models: (E) 1-mers+2-mers+3-mers on all (1,032) sequences (not only unique flanking sequences as in Figure 4). (F) 2 bp immediate flanks, 1-mers+2-mers on 103 sequences. (G) 1 bp immediate flanks, 1-mer 412 sequences. (H) 3 bp flanks at positions 4-6 on the 5' and 3' sides, separated by 3 bp from the core TFBS, 1-mers+2-mers+3-mers, 412 sequences.



A	0.249	0.222	0.187	1	0	0	1	0	0	0	1	0	0.283	0.226	0.207
C	0.254	0.246	0.181	0	0	0	0	1	0	1	0	0	0.297	0.240	0.221
G	0.227	0.242	0.272	0	0	1	0	0	0	0	0	0	0.159	0.236	0.265
T	0.269	0.290	0.360	0	1	0	0	0	1	0	0	1	0.261	0.298	0.307

A	0.226	0.268	0.159	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0.452	0.353	0.301
C	0.277	0.224	0.297	1	0	0	0	0	0	0	1	0	1	0	1	1	0	1	1	0	0.145	0.214	0.305
G	0.325	0.241	0.157	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0.266	0.255	0.238
T	0.173	0.267	0.387	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0.137	0.178	0.156

Figure S7. An alternative representation, in the form of a PWM, of the flanking sequences preferences observed in our measurements. (A-B) A Position Weight Matrix was derived in the following manner: For nucleotide *i* in position *j* we computed the sum of binding scores of sequences that had this nucleotide in this position. This sum is then normalized by the number of such sequences; so that entry *i, j* in the matrix in fact holds the average score of sequences that had nucleotide *i* in position *j*. These values were then transformed to frequencies, so that every potion sums up to one. (A) A Position Weight Matrix derived from 412 sequences with unique 15bp flanks surrounding a strong 9-bp Gcn4 core sites (the same set of sequences for which the model in figure 4B was constructed). Preferences captured by the linear model (see Figure 4A,D) can be observed in the PWM. (B) A Position Weight Matrix derived from 315 sequences with unique 15bp flanks surrounding a strong 17-bp Gal4 core sites (the same set of sequences for which the model in figure 4C was constructed). Preferences captured by the linear model (see Figure 4E) can be observed in the PWM. (C-D) A score per sequence was computed by multiplying the relevant frequencies at each position. 3-bp flanks were used for prediction, and comparison to the linear models performance. (C) Shown is the score computed based on the PWN in A versus the measured Gcn4 binding score (with Pearson's correlation of 0.73, compared to 0.86 obtained with the linear model, see figure 4B). (D) Shown is the score computed based on the PWN in A versus the measured Gal4 binding score (with Pearson's correlation of 0.819, compared to 0.95 obtained with the linear model, see figure 4C).

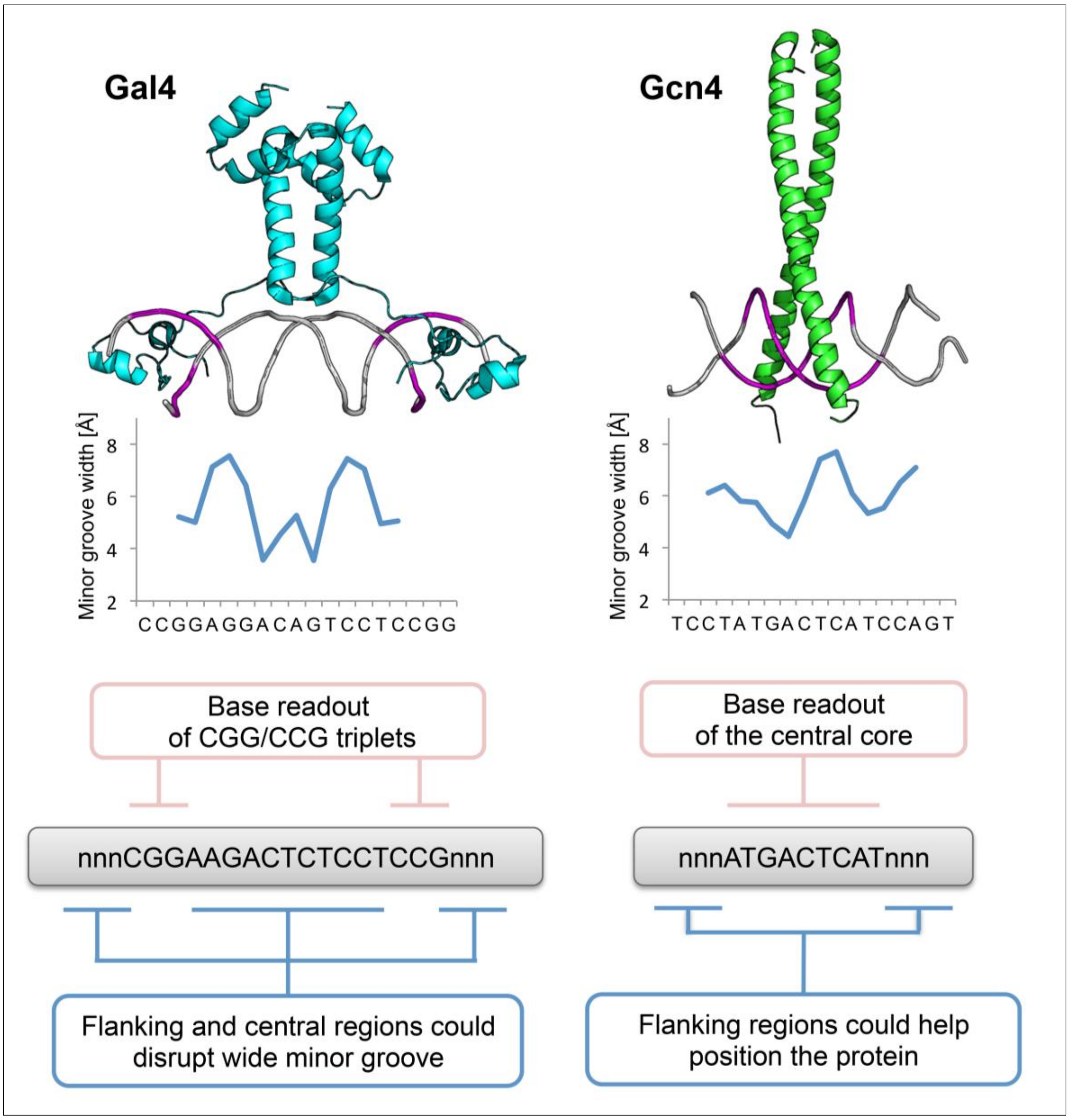

Gal4
Gcn4
Minor groove width [Å]
8
6
4
2
CCGGAGGACAGTCCTCCGG
Minor groove width [Å]
8
6
4
2
TCCTATGACTCATCCAGT
Base readout of CGG/CCG triplets
Base readout of the central core
nnnCGGAAGACTCTCCTCCGnnn
nnnATGACTCATnnn
Flanking and central regions could disrupt wide minor groove
Flanking regions could help position the protein

Figure S8. Schematic illustrations of the distinct DNA recognition mechanisms used by Gal4 and Gcn4. Gal4 binds as a homodimer (PDB ID 3COQ) to a 17-bp site, with only 3 bp at the 5′ and 3′ ends of this binding site in direct contact with the protein (base readout; pink box), which are therefore highly conserved. The central 11 bp between the CGG/CCG triplets are not directly contacted and, thus, not conserved, which explains the differences in some nucleotide positions between the co-crystal structure (MGW plot) and the BunDLE-seq data (gray box). Flanking sequences and the variable region between the CGG/CCG triplets can have a larger impact on the relatively variable Gal4 site, for instance by disrupting the width of the minor groove (shape readout; blue box). Gcn4 binds DNA as a bZIP homodimer (PDB

ID 1YSA) mainly through an intensive network of hydrogen bonds with the major groove edges of the central 7 bp of its binding site (base readout; pink box), which are therefore highly conserved. Flanking sequences can contribute to the fine-tuning of the binding specificity of Gcn4 to the core-binding site (shape readout; blue box).

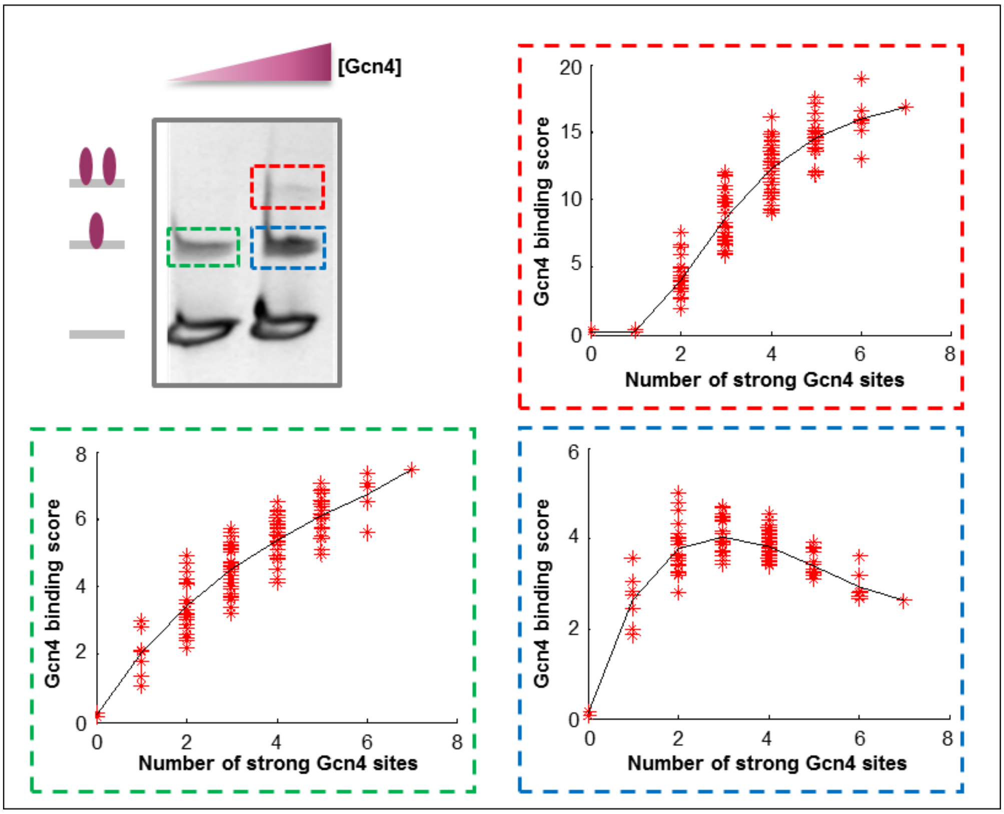


Figure S9. Dependency of different binding events on the number of TFBS sites. For a set of sequences with all possible combinations of one to seven binding sites for Gcn4, in seven predefined locations, plotted is the binding score in different bands (‘binding states’), as a function of the number of sites within each sequence (in red), as well as the mean binding score across the sequences sharing the same number of sites (in black). The graphs correspond to the bands displayed in the gel (as follows from the color of the square surrounding the band/graph). Note that for the band marked in red (that is the second band to appear when the TF is present), sequences with either zero or one binding sites extremely rare, supporting the notion that this band represents DNA molecules bound by two TF molecules.

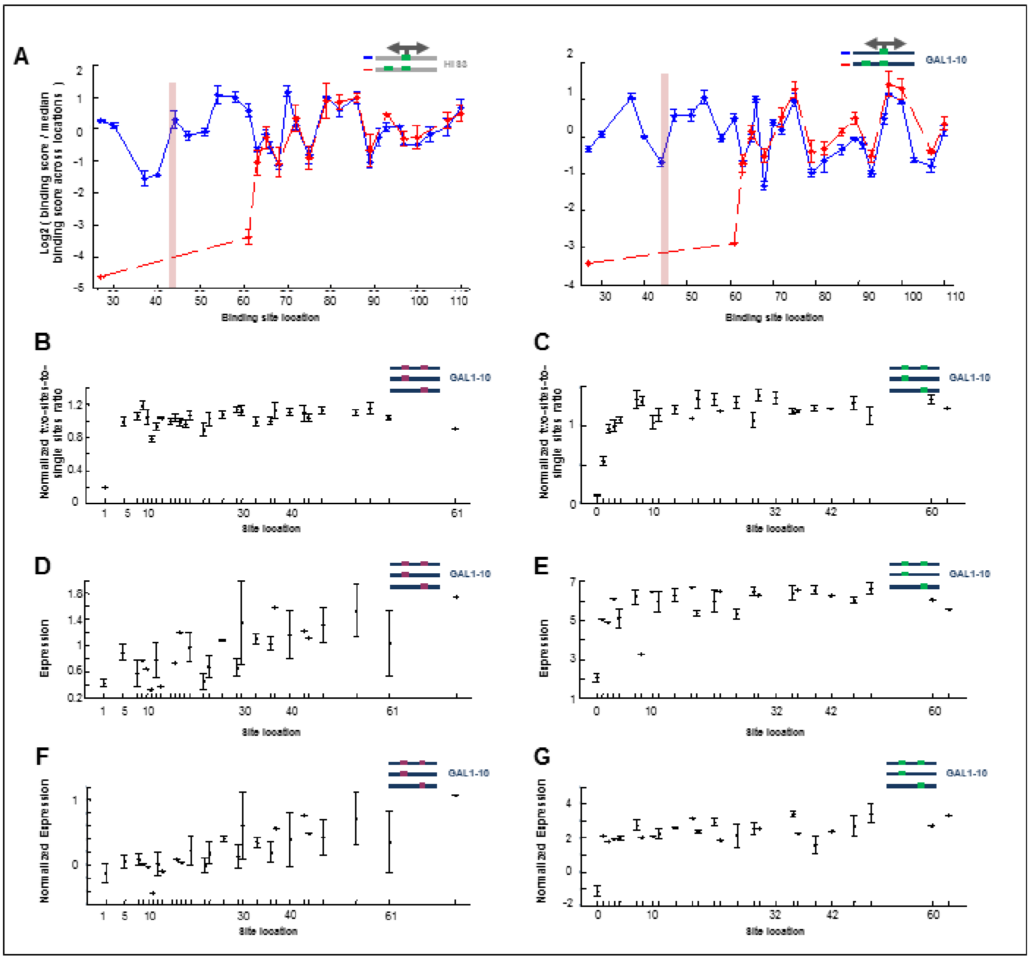


Figure S10. Closely located TFBSs can influence, for instance in an inhibitory manner, co-occurring TF binding to the sites. (A) Same as the graphs in Figure 5A, with a Gal4 TFBS of 17 bp placed along the *HIS3*-derived context (left panel) and along the *GAL1-10*-derived context (right panel). In the two-site sequences that deviate from the one-site sequences pattern, sites were completely adjacent. (B,C) For a set of sequences with two TFBS plotted is the mean +/- std of the “normalized two sites-to-single site ratio” measure as a function of the distance between the edge of one site to the edge of the other (i.e. the number of bps separating the two sites). The “normalized two sites-to-single sites ratio” measure is computed by dividing the binding score of the sequence containing two sites, computed based on a band representing two TF binding, by the product of the binding scores of the two corresponding sequences containing one of the sites, that are computed based on the band representing a single TF binding.

Notably, the division by the product of binding scores of the corresponding single site sequences accounts for the effect of the specific sites locations.

In fact this computed ratio is proportional to what can be considered a cooperatively factor (a cooperatively factor multiplied by the constant *C*):

$$\frac{p(site1 = 1\ \cap site2 = 1)}{p(site1 = 1) * p(site2 = 1)} * C$$

With the constant $C = \frac{p(1\mathrm{-}TF\mathrm{-}bound\mathrm{-}band)^2}{p(2\mathrm{-}TF\mathrm{-}bound\mathrm{-}band)}$

Mostly for visualization purposes, we further divided this ratio by a normalizing factor. This factor is the mean binding score for sequences with two sites in the two TF binding band divided by the square of the mean binding score of sequences with a single score in the one TF binding band, as computed across a controlled subset of sequences in which the sites are placed in all possible locations out of seven predefined locations, as can be seen in Figure S6. This normalization factor is computed for a controlled subset of sequences and averages over different distances and found to bit successfully fitted by a thermodynamic model assuming independence multiple TF binding events, it is thus reasonable to assume the cooperatively factor in this case will be close to 1 and the value of this normalization factor will be close to C. So that once we divide by this normalization factor we cancel out C and get the measure to be closer to the cooperatively factor. Indeed for most distances the values we get after this normalization are close to 1, while for the closely located sites the value is below 1, likely indicating an inhibitory effect of one binding event on the other. This matches the observation in Figure 6A and panel A in this figure. (B) This measure is plotted for sequences with a Gcn4 9 bp sites, placed along the *GAL1-10*-derived context. (C) This measure is plotted for sequences with a Gal4 17 bp sites, placed along the *GAL1-10*-derived context. (D) For the same sequences presented in B (with two TFBS for Gcn4 placed along the *GAL1-10*-derived context), shown is the mean +/- std of the expression measurements for the two sites sequences as a function of the distance between the edge of one site to the edge of the other (i.e. the number of bps separating the two sites). (E) For the same sequences presented in C (with two TFBS for Gal4 placed along the *GAL1-10*-derived context), shown is the mean +/- std of the expression measurements for the two sites sequences as a function of the distance between the edge of one site to the edge of the other (i.e. the number of bps separating the two sites). (F) For the same sequences presented in B (with two TFBS for Gcn4 placed along the *GAL1-10*-derived context), shown is the mean +/- std of a normalized expression measure as a function of the distance between the edge of one site to the edge of the other (i.e. the number of bps separating the two sites). This measure is simply the difference between the expression level of the two sites sequence and the sum of expression levels of the corresponding sequences single site sequences. (G) For the same sequences presented in C (with two TFBS for Gal4 placed along the *GAL1-10*-derived context), shown is the mean +/- std of a normalized expression measure as a function of the distance between the edge of one site to the edge of the other (i.e. the number of bps separating the two sites). This measure is simply the difference between the expression level of the two sites sequence and the sum of expression levels of the corresponding sequences single site sequences.

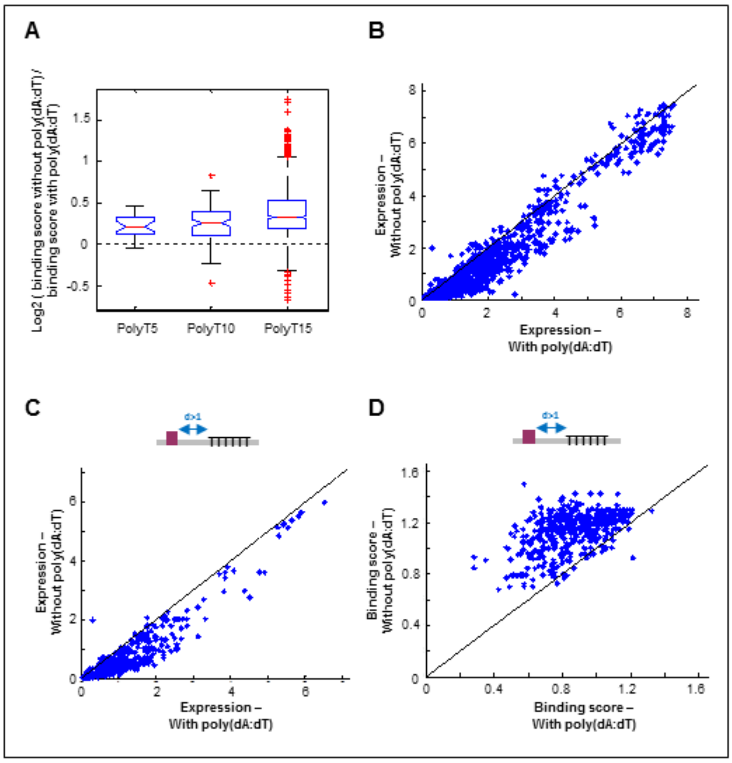


Figure S11. Nucleosome sequence preferences can be captured by BunDLE-seq and thereby suggest additional mechanisms underlying expression differences between different regulatory sequences. (A) From left to right: Shown is a boxplot of the log2 ratio of the nucleosome binding score for ~20 sequences with a poly(dA:dT) tract of length 5 bp and the corresponding sequences lacking this tract, same for ~100 sequences with a poly(dA:dT) tract of length 10, same for ~2000 sequences with a poly(dA:dT) tract of length 15 bp. (B) Scatter plots of expression measurements for the same sequences as in Figure 6C, that is with or without a 15 bp poly(dA:dT) tract. (C) Scatter plots of expression measurements for the same sequences as in Figure 6C, but for sequences in which the 15 bp poly(dA:dT) tract is located more than 1 bp away from the binding site. (D) Scatter plots of nucleosome binding scores for the same sequences as in B.

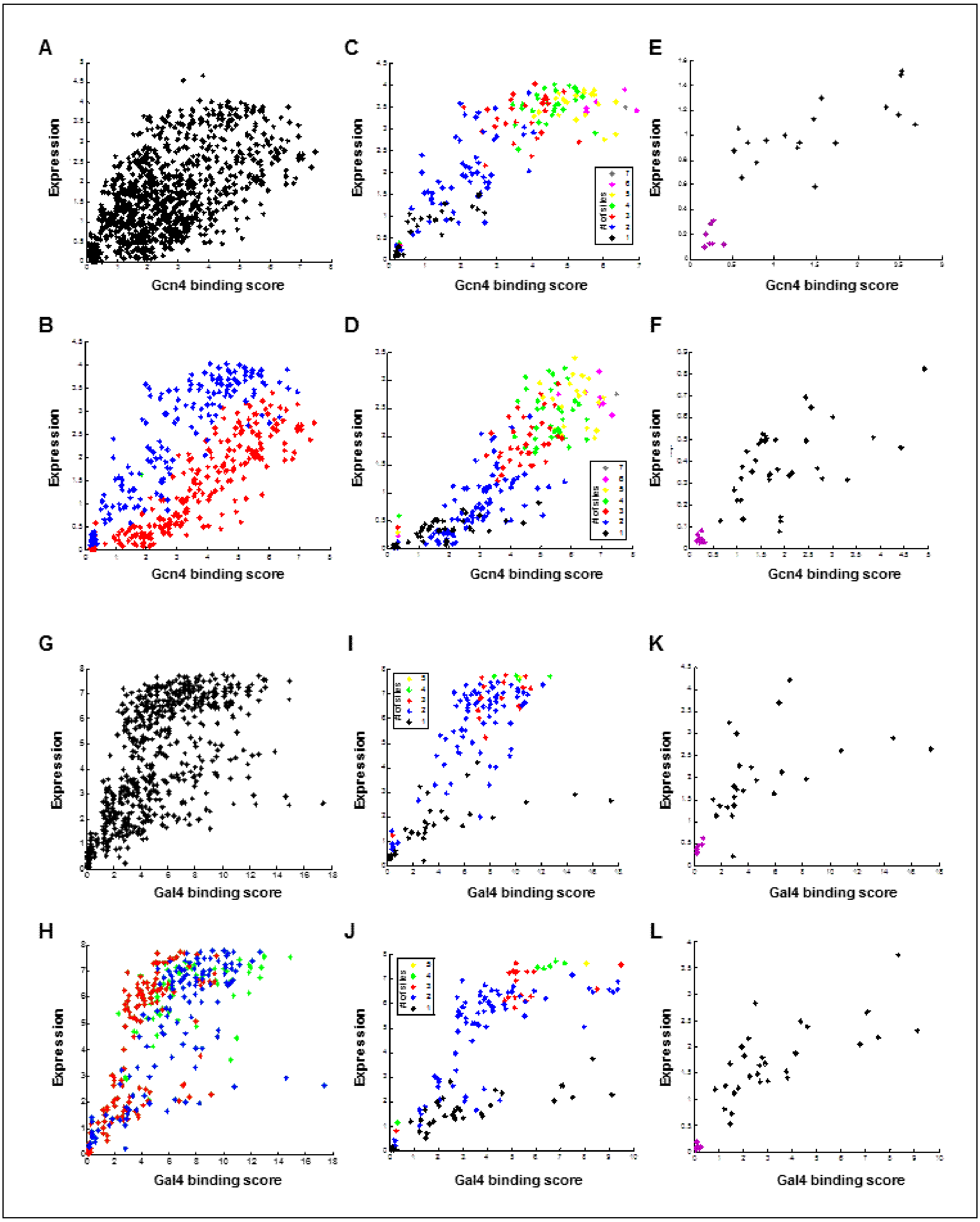


Figure S12. A comparison of BunDLE-seq binding measurements and corresponding expression measurements. Shown is the binding score (obtained for a single TF binding event) versus the expression measurements obtained with same sequences serving as promoters in high-throughput reporter activity assay carried out in yeast cells; for the following sets of sequences: (A) All sequences with a Gcn4 TFBS, and no other designed

TFBSs (1050 sequences). (B) All sequence with a Gcn4 TFBSs, but no poly(dA:dT) tracts (418 sequences). Sequences with *HIS3*-derived context are colored in blue and sequences with *GAL1-10*-derived context colored in red. (C) Sequences with a Gcn4 TFBS, placed along the *HIS3*-derived context (same as the sequences colored in blue in B), color-coded according to the number of TF sites in the sequence (172 sequences). (D) Sequences with a Gcn4 TFBS, placed along the *GAL1-10*-derived context (same as the sequences colored in red in B), color-coded according to the number of TF sites in the sequence (245 sequences). (E) Sequences with only a single GCn4 site, placed along the *HIS3*-derived context (same as the sequences colored in black in C) (25 sequences). (F) Sequences with only a single GCn4 site, placed along the *GAL1-10*-derived context (same as the sequences colored in black in D) (48 sequences). (G-L) same as A-F, but with Gal4 sites, instead of Gcn4 sites (with 663, 332, 130, 141, 33, 41 sequences in G-L, accordingly).

References

LeProust EM, Peck BJ, Spirin K, McCuen HB, Moore B, Namsaraev E, Caruthers MH. 2010. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic acids research* **38**(8): 2522-2540.

Quail MA, Swerdlow H, Turner DJ. 2009. Improved protocols for the illumina genome analyzer sequencing system. *Current protocols in human genetics / editorial board, Jonathan L Haines [et al]* **Chapter 18**: Unit 18 12.

Raveh-Sadka T, Levo M, Segal E. 2009. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome research* **19**(8): 1480-1496.

Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, Zeevi D, Sharon E, Weinberger A, Segal E. 2012. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nature genetics* **44**(7): 743-750.

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature biotechnology* **30**(6): 521-530.

Struhl K, Segal E. 2013. Determinants of nucleosome positioning. *Nature structural & molecular biology* **20**(3): 267-273.

Zhao Y, Granas D, Stormo GD. 2009. Inferring binding energies from selected binding sites. *PLoS computational biology* **5**(12): e1000590.