Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays

Graphical Abstract

Highlights
- Parallel expression and occupancy measurements reveal TF-nucleosome interplay
- Pronounced differences in transcription factors (TFs) sensitivity to chromatin
- Specific combinations of TF sites give rise to nucleosome-mediated collaborations
- Properties of nucleosome-based collaborations can quantitatively shape expression

Authors
Michal Levo, Tali Avnit-Sagi, Maya Lotan-Pompan, Yael Kalma, Adina Weinberger, Zohar Yakhini, Eran Segal

Correspondence
eran.segal@weizmann.ac.il

In Brief
With parallel measurements of gene expression and DNA occupancy, Levo et al. examine transcription factors (TFs) activity in the context of chromatin. Regulatory sequences, with various combinations and arrangements of TF sites, give rise to different nucleosome-mediated collaborations between TFs. These, in turn, can quantitatively account for the resulting expression.
Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays

Michal Levo, Tali Avnit-Sagi, Maya Lotan-Pompan, Yael Kalma, Adina Weinberger, Zohar Yakhini, and Eran Segal

1Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel
2Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel
3Computer Science Department, Technion, Haifa 3200003, Israel
4Agilent Laboratories, Santa Clara, CA 95051, USA
5Present address: Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
6Lead Contact
*Correspondence: eran.segal@weizmann.ac.il
http://dx.doi.org/10.1016/j.molcel.2017.01.007

SUMMARY

Precise gene expression patterns are established by transcription factor (TFs) binding to regulatory sequences. While these events occur in the context of chromatin, our understanding of how TF-nucleosome interplay affects gene expression is highly limited. Here, we present an assay for high-resolution measurements of both DNA occupancy and gene expression on large-scale libraries of systematically designed regulatory sequences. Our assay reveals occupancy patterns at the single-cell level. It provides an accurate quantification of the fraction of the population bound by a nucleosome and captures distinct, even adjacent, TF binding events. By applying this assay to over 1,500 promoter variants in yeast, we reveal pronounced differences in the dependency of TF activity on chromatin and classify TFs by their differential capacity to alter chromatin and promote expression. We further demonstrate how different regulatory sequences give rise to nucleosome-mediated TF collaborations that quantitatively account for the resulting expression.

INTRODUCTION

Different regulatory DNA sequences can drive distinct levels of gene expression. Large collections of gene expression measurements, as well as transcription factor (TF) binding and chromatin occupancy measurements are now available. Yet, a systematic understanding of how these properties are related to one another and how they depend on the DNA sequence is still missing (Levo and Segal, 2014). This limits our ability to interpret the non-coding regions of the human genome and assess the relevance of individual’s sequence variation to gene expression and ultimately incidence of disease.

In the past decade, high-throughput in vitro TF binding measurements significantly improved our ability to identify motifs within regulatory sequences, yet these often do not match measured binding events within cells (Levo and Segal, 2014). In turn, TF binding in cells is often not predictive of an expression outcome (Spitz and Furlong, 2012; White et al., 2013). Identifying and characterizing the mechanisms that account for these pronounced differences remains a challenge. TFs competition with nucleosomes and the combinatorial nature by which TFs bind to regulatory sequences were suggested over the years to play a significant role (Levo and Segal, 2014; Spitz and Furlong, 2012). Yet, it is less clear how the functionality of different motifs depends on the surrounding chromatin environment and how different regulators, whose motifs are clustered together, act in concert.

High-throughput quantitative reporter assays offer an appealing strategy to address the above questions, as these assays provide means to measure the effects of carefully designed manipulations to regulatory sequences on the resulting expression (Inoue and Ahtuv, 2015; Levo and Segal, 2014). Yet, currently we lack an approach that allows high-resolution quantitative measurements of DNA binding that can complement these expression measurements. As a result, the mechanisms underlying the measured differences in expression often remain unclear, making it difficult to generalize the observed effects and predict the consequences of additional sequence manipulations.

Here, we present BE-MPRA, a binding and expression massively parallel reporter assay, for accurate measurements of DNA occupancy in large-scale reporter libraries. We demonstrate the application of this assay to a systematically designed library of >1,500 promoter variants in yeast. In this assay, we utilize an approach previously used to interrogate several specific loci (Fatemi et al., 2005; Gal-Yam et al., 2006; Jessen et al., 2006; Kilgore et al., 2007; Small et al., 2014; Stees et al., 2015) or to measure mono-nucleosomes occupancy (as well as DNA methylation) genome-wide (Kelly et al., 2012; Taberlay et al., 2014). The method makes use of a methyltransferase that methylates accessible DNA, followed by bisulfite conversion.
“recording” the methylation status (indicative of the DNA occupancy) into the sequence.

By employing this approach, our assay reveals, with high-resolution, differences in the occupancy and organization of promoter’s nucleosomes, as well as captures for some TFs, adjacent yet distinct binding events. Importantly, the assay provides means to quantify the fraction of cells in which a region is occupied and to examine the occupancy pattern at the single-cell level.

As the assay is tailored for large-scale reporter libraries, occupancy measurements are readily coupled with expression measurements, shedding light on how TF-nucleosomes interplay quantitatively shapes the expression outcome. We find that the ability of a TF to promote expression can differ significantly when its cognate binding site is embedded in an open versus a closed chromatin environment. We further show that TFs differ in this sensitivity to the chromatin environment. With our assay, we measure the differential capacity of TFs to alter chromatin or promote expression on their own. We then characterize how different regulatory sequences give rise to collaborations between TFs that differ in these capacities. Finally, we demonstrate the utility of our approach in dissecting the logic of native promoter sequences.

RESULTS

High-Resolution In Vivo Binding Measurements on Large-Scale Libraries of Designed Promoter Sequences

We designed a library of >1,600 promoter sequences (of length 210 bp) that were then synthesized on Agilent programmable microarrays (LeProust et al., 2010). We genomically integrated the variants into a fixed location in the yeast genome upstream a yellow fluorescence protein (YFP) reporter gene. We grew the obtained population of yeast cells to mid-log under conditions activating the main regulators used in our design (i.e., galactose and amino-acid starvation activating Gal4 and Gcn4). We then performed either promoter activity measurements (hereinafter referred as expression measurements) as previously described (Sharon et al., 2012) or applied the developed assay for high-resolution DNA occupancy measurements.

In this assay, we utilize the differential activity of a DNA methyltransferase in the context of accessible versus occluded DNA (Kladde et al., 1996). Specifically, a methyltransferase (M.SSsI) that we introduce to crosslinked yeast spheroplasts carrying our library, methylates accessible occurrences of CpG. We then extract the DNA and perform bisulfite conversion (Clark et al., 1994), converting only unmethylated cytosines to thymines. We amplify 520 bp surrounding our designed promoter variants and read out the resulting sequences with high-throughput, 600 bp long, MiSeq reads. An original CpG cytosine that remained a cytosine is indicative of methylation and is hence suggestive of accessibility, while a cytosine that was converted to a thymine indicates lack of methylation and hence occlusion by some DNA binding protein (Figures 1A and 1B). We map each read back to a designed variant in our library, and per designed variant, we use the percentage of reads in which each CpG was unmethylated (converted) to compute a population-averaged occupancy pattern (Figure 1C). Notably, in contrast to chromatin immunoprecipitation (ChiP)- or micrococcal nuclease (MNase)-based assays where only occupied fragments are collected, our method reports both on occupied and unoccupied fragments. Thereby, this assay divulges the fraction of the cell population in which a region was occupied (see the STAR Methods). Furthermore, the method does not involve any selection of occupied fragments based on the identity of the binding molecule (as in ChiP assays) or the size of the occupied region (as in MNase-based nucleosome mapping), and it is therefore able to report on both TF and nucleosome binding events.

We note that while nucleosomes are readily captured, capturing TF binding events is more challenging and depends heavily on the TF identity and its binding site (as these can affect how the TF structurally occludes the methyltransferase). Importantly, captured binding events are measured with remarkable resolution, as evident, for example, by three distinct footprints over three adjacent binding sites for the TF Abf1, occurring within a region of <100 bp (Figures 1B–1D). Moreover, as the assay does not involve fragmentation of the DNA in the region of interest, adjacent binding events co-occurring within a single cell are registered within the same sequencing read (Figure 1B). The use of long-read sequencing thus enables us to read out this single-cell information. This provides the unique opportunity to capture the diverse single-cell DNA binding configurations formed on different regulatory sequences and assess how cells distribute among them (Figure 1D).

We find that our binding measurements are highly reproducible as replicates exhibit very similar occupancy patterns (r = 0.86). To further validate the quantitative and reproducible nature of our assay, we included in our designed library sets of sequence variants differing only by a 10 bp barcode and found that these sequences show highly similar occupancy patterns (r = 0.86, Figures S1A and S1B). Notably, these sets of sequences also allow us to assess the reproducibility of our expression measurements, and indeed, we find that sequences differing only in their barcode show highly similar expression levels (Figures S1C and S1D).

A Dual Role for Transcription Factors

Our experimental setup provides means to examine the effect of the chromatin environment in which TF binding sites (TFBSs) are embedded on the activity of their respective regulators. To this end, we selected 12 yeast TFs and designed promoter variants in which a single site is introduced in a fixed location, −135 bp upstream of the transcription start site (TSS) (see Table S1 for site’s sequences and Figures S2A and S2B for a comparison to variants with mutated sites). All sites were introduced within two sequence contexts (Figures 2A and 2B). One sequence context (derived from the native HIS3 promoter) displays a largely nucleosome-free region (NFR) around −200, flanked by the −1 and +1 nucleosome (herein referred as the open chromatin context). The other sequence context (derived from the native GAL1-10 promoter) has three nucleosomes occupying the promoter, i.e., a nucleosome occupies the region that is depleted in the open chromatin context (hereinafter referred to as the closed chromatin context). For simplicity and ease of comparison between contexts, we will refer to this middle nucleosome as an NFR-nucleosome and keep referring to the...
We find that different TFBSs introduced into the open chromatin context (Figure 2A) differ in their capacity to drive expression, with the Gal4 site, for example, driving the highest expression. Such differences can stem from differential affinity of the sites, differential concentration of the corresponding regulator, or differential capacity of the regulators to promote expression. Interestingly, we find that some TFBSs, for instance an Abf1 site, have little or no effect on expression even though a footprint over the site indicates binding of the TF.
Figure 2. TFs Differ in Their Ability to Alter Chromatin or Promote Expression
(A) TFBSs for different TFs were placed (centered −135 bp upstream of the TSS) in a sequence derived from the HIS3 native promoter showing an open chromatin landscape. Occupancy pattern prior to sites additions is shown both as an area plot and as a heatmap. Below: the occupancy pattern for a variant with each TFBS (rows of the heatmap, TFBS illustrated on the left) and the measured expression (bar plot) for that variant is shown.
(B) Same as (A) with TFBSs placed in a sequence derived from the GAL1-10 native promoter showing a closed chromatin landscape.
(C) For each TFBS, the log2 ratio of expression measurements of the variant in the open chromatin context and the corresponding variant with in the closed chromatin context is shown. TFBS are sorted by this log2 ratio with the color-coded rectangles at the bottom of each bar indicating the TFBS identity.
(D) For each TFBS, the mean difference in occupancy per base pair in the window marked in a black rectangle in (B) is shown (between −50 to −200 bp upstream of the TSS, excluding 25 bps surrounding the center of the added site), between the variant in which the TFBS was placed in the closed chromatin context and the variant lacking this TFBS (i.e., the corresponding row in B relative to the bottom row in B).
(E) A scatterplot of the expression shown in (B) versus the difference in occupancy in the site vicinity upon site addition shown in (D). Dots are color-coded by the identity of the added TF. Pearson correlation across all examined TFs is $r = -0.63$. This correlation increases to $r = -0.93$ when excluding Rap1, Abf1, Reb1, Sfp1, and Hap1 (TFs whose addition results in significant change in occupancy yet a minor change to expression).

When we introduce the same TFBS, in the same position, into the closed chromatin context, we find expression is always lower than that obtained by the corresponding open chromatin variant (Figures 2B and 2C). Moreover, the ranking of these single-site variants by their expression is different in these two contexts (Figures 2B and 2C). Our occupancy measurements suggest some insights as to the source of these differences. For some of our examined TFs, introducing a site into the closed chromatin context results in a significant reduction in expression in the site vicinity. Namely, we observe >23% reduction for Sfp1, Reb1, Abf1, Rap1, and Gal4 sites (Figure 2D). This is consistent with previous studies associating these TFs with chromatin remodeling capabilities (Ganapathi et al., 2011; Henikoff et al., 2011; Wang et al., 2011). Interestingly, despite this pronounced change in occupancy, with the exception of Gal4, introducing these sites does not increase expression significantly (Figures 2B and 2E). For our other examined TFs, a single site introduced into the closed chromatin context seems incapable of significantly altering chromatin. We observe <3.5% difference in occupancy in the site vicinity for Bas1&2, Gcn4, Fhl1, Gcr1, or Hap2–Hap5 sites (Figure 2D). The promoter remains largely occupied with >80% of cells displaying an NFR-nucleosome. Correspondingly, the expression of these variants is relatively low (Figures 2B and 2E). It is this group of TFs that show the largest differences in expression when sites are embedded in the open versus the closed chromatin context (Figure 2C); most likely as the inability to alter chromatin renders them more susceptible to the chromatin environment.

Our assay thus reveals pronounced differences between TFs in their sensitivity to the chromatin environment in which a motif is embedded and their ability to alter this environment. While some TFs are able to alter chromatin and drive strong expression regardless of the initial chromatin landscape (e.g., Gal4), others can alter chromatin but do not seem to contribute significantly to expression on their own (e.g., Abf1). A third group of TFs lacks the ability to alter chromatin (at least with a single site) yet can directly contribute to expression if their cognate site is accessible for binding (e.g., Gcn4).
Homotypic Clustering in the Context of Chromatin

Regulatory sequences, from yeast to higher eukaryotes, are often characterized by the presence of multiple adjacent TFBSs (Spitz and Furlong, 2012). It was suggested that such TFBS clusters facilitate the cooperation between TFs in their competition with nucleosomes for DNA binding (Miller and Widom, 2003; Mimny, 2010; Poich and Widom, 1996; Vashee et al., 1998). Reporter assays offer means to systematically examine such clusters (Fiore and Cohen, 2016; Sharon et al., 2012; Smith et al., 2013). Yet the scarcity of complementary in vivo occupancy measurements limited our mechanistic understanding of how expression depends on the site’s identity, multiplicity, and sequence context.

We thus sought to use our assay to examine TFBS clustering starting with homotypic clusters (i.e., multiple sites for the same TF). For several TFs, we introduced all $2^k$ possible combinations of 0 up to k TFBSs at predefined locations in both the open and closed chromatin contexts (>340 variants, Figures 3 C and S2C–S2H). Previous studies showed that expression often increases monotonically with the number of binding sites (Sharon et al., 2012), although with different curves being reported (Levo and Segal, 2014). Here too, we find the maximal expression level and the shape of the curves describing the dependency of expression on the number of sites (averaging across different site locations) differs not only between TFs, but also between the two examined contexts. Open chromatin variants show higher expression than the corresponding closed chromatin variants. Moreover, the rise in expression in the curves obtained in the closed chromatin context is delayed (although often subsequently sharper) compared to that observed in the open chromatin context (Figures 3 C–3F and S2C–S2H). To test whether nucleosome occupancy can indeed underlie these differences, we computed the average occupancy pattern along the promoter for variants with the same number of sites. We focused on Gcn4 and Gcr1 as our above analysis demonstrated a single site for these TFs is insufficient to alter chromatin, suggesting binding would be susceptible to chromatin, and site multiplicity could be beneficial (Chambers et al., 1998; Uemura et al., 1997; Zeigler and Cohen, 2014). Indeed, we find the mean occupancy (Figures 3 C–3F) shows high agreement with the measured mean expression ($r = −0.87$, p value $<0.0003$ for Gcn4 variants, and $r = −0.98$, p value $<10^{-9}$ for Gcr1 variants). Specifically, in the closed chromatin context, we find that in the presence of a single Gcn4 site, or even three Gcr1 sites, a pronounced nucleosome is observed between the +1 and −1 nucleosomes, corresponding to the attenuated rise in expression. The addition of more TFBSs facilitates the eviction of this nucleosome and the sharper rise in expression. As we add sites to the open chromatin context, the NFR observed, even prior to the addition of any TFBSs, is further extended, with a downstream shift of the +1 nucleosome and weakening of the −1 nucleosome, and expression rises steadily.

Thus, by coupling expression and occupancy measurements, we find evidence for nucleosome-mediated cooperativity in the context of homotypic TFBS clustering. We demonstrate that different TFs require a different number of sites in order to elicit nucleosome eviction. We show how this thereby accounts for differences in the quantitative dependency of expression on site multiplicity between TFs and between contexts.

Heterotypic Clustering in the Context of Chromatin

The common occurrence of heterotypic clustering of TFBSs in native yeast promoters and higher eukaryotes enhancers often brings together a TF capable of eliciting some change to chromatin (e.g., pioneer factors) (Zaret and Mango, 2016) and TFs that seem to lack this ability, yet likely benefit from the activity of the first (Ganapathi et al., 2011; Spitz and Furlong, 2012). However, very little is known on the properties of such cooperation.

To address this question, we designed >230 promoter variants with combinations of potentially cooperating TFs. An initial set of variants was based on native yeast promoters (HIS3, PGK1, and ENO1) that include annotated elements suspected to affect chromatin, namely Poly(dA:dT) tracts (Iyer and Struhl, 1995; Raveh-Sadka et al., 2012) known to be nucleosome disfavoring (Struhl and Segal, 2013), and binding sites for Abf1, Rap1, and Reb1 (Chambers et al., 1995, 1988; Uemura et al., 1997). When mutating these elements, we found an increase in nucleosome occupancy and a reduction in expression (Figure 4A). To better understand how these elements differ in their effect, we further replaced the native element with a consensus site for either Gal4, Rap1, Abf1, Reb1, or a Poly(dA:dT) tract (Figure 4A). We also placed each of these chromatin-altering elements within our two synthetic contexts in two locations and in proximity to sites for other TFs (Figures 4B and S3). We find that all of these elements contribute to reduced nucleosome occupancy in the promoter region, although they differ in their exact effect (Figures 5 and S3). While the addition of a Gal4 site generally results in a reduction in all of the promoter’s nucleosomes (consistent with Wang et al., 2011), a Rap1 site results in reduced occupancy in the NFR, often a downstream shift in the +1 nucleosome and a reduced occupancy of the −1 nucleosome. The addition of an Abf1 site, a Poly(dA:dT) tract, and a Reb1 site often results in a more local reduction in occupancy, mostly downstream of the sites (consistent with Yarragudi et al., 2004). Indeed, placing these elements in a more upstream position resulted in an upstream extension of the formed NFR (Figure S3). Notably, the nature of the effect conferred by each specific chromatin-altering element is similar across our examined contexts, although the magnitude of the effect can differ (Figures 5A, 5B, and S3). For instance, the presence of a nearby TFBS can further strengthen the reduction in occupancy conferred by the chromatin-altering element (e.g., occupancy decreases 2.5 times more when a Rap1 or an Abf1 site is added near a Bas1&82 site than when it is added alone, Figure 5B). When we quantify the reduction in occupancy upon the addition of the chromatin-altering elements, we further find it to be in high alignment with the measured increase in expression (Figures 5B and S4; Pearson correlations ranging from $−0.81$ to $−0.99$ in the different contexts, excluding those that had the Gal4 site and were therefore less influenced by this addition).

Indeed, corresponding to the differential reduction in occupancy in the presence of different nearby TFs, we observe a differential benefit in terms of expression increase (Figure S4F). More generally, we find that the modulation of expression by different nearby TFBSs is similar across different chromatin-altering elements (mean Pearson correlation $r = 0.89$ between the different color-coded lines in Figure S4G). It thus seems
that expression can be increased in a modular manner by utilizing a TF that is more potent in altering chromatin and a nearby TF that is more potent in promoting expression.

Based on these principles, we devised a simple model to predict the expression of variants with two TFBSs based on our occupancy and expression measurements for the corresponding single-site variants (see Figure S5 for a detailed description). If the two sites were embedded in the open chromatin context, our model adds the expression contribution of each of these sites. If the two sites were embedded in the closed chromatin context, our model first assesses the fraction of the population that has open chromatin (i.e., an expressing state) and the fraction of the population that has closed chromatin (i.e., lowly expressing state). These fractions are computed based on the reduction in occupancy conferred by the composing chromatin-altering element. For the fraction of the population with an open chromatin state, the model then adds the expression contributions of the composing sites (whereas the closed chromatin state only contributes a basal expression level). As an example, in this model, a Rap1 site would result in a larger fraction of the population in the open chromatin state, compared to an Abf1 site. The presence of a nearby Bas1&2 site would further result in a higher expression from the open state than the presence of an Fhl1 site. We find that this simplified model accounts well for our measurements, with an $R^2 = 0.87$, and provides a plausible mechanistic intuition as to the roles of the composing TFBSs.

To gain further insight as to the determinants of the cooperation between a chromatin-altering element and the nearby benefitting TF, we examine another set of variants. In these, we focus on a site for Gcn4 and alter more systematically the strength of the chromatin-altering element and the location of both elements within the closed chromatin context. Consistent with our pervious study (Raveh-Sadka et al., 2012), we find that...
we can increase expression by employing a longer poly(dA:dT) tract or a shorter distance between the tract or a site for Abf1 or Rap1 and the nearby Gcn4 site. Our occupancy measurements reveal that a “stronger” element, and even more so, a more upstream position of the chromatin-altering element results in an upstream extension of the NFR (Figures 6A and 6B). Yet, it is not this extension of the NFR that correlates with the increase in expression, as can be seen from the higher expression of the variants in which the element was placed in a more downstream position, closer to the Gcn4 site and the TSS. By calculating mean occupancy along the promoter with a running window, we find that the region with highest anti-correlation between expression and occupancy is the region containing the Gcn4 site and the base pairs downstream to it ($r = -0.91$, p value $< 10^{-7}$, Figure 6C).

We further find that even if we maintain a constant distance between the two composing elements, we can modulate the expression contribution of the chromatin-altering element by changing the location of these elements within the promoter and relative to one another. When the Gcn4 site is placed
close to the midpoint of the NFR nucleosome (−195 bp), it drives a low expression and benefits more from the addition of a chromatin-altering element (placed at −225 bp) than when these sites are switched in their locations (Figures S6 A and S6B).

Our assay thus reveals that different chromatin-altering elements possess a distinct, prototypical effect on chromatin that is maintained across contexts. The quantitative implications of these chromatin alterations on expression depends on the initial chromatin landscape, the identity of the benefiting TF, and the placement of these elements with respect to the promoters’ nucleosomes and to one another. We find, for example, that it is not the upstream expansion of the NFR but rather the local accessibility in the region of the benefiting TF site and possibly the downstream expansion of the NFR, that are positively correlated with expression.
Unraveling the Complexity of Native Sequences

We next examined seven well-annotated native yeast promoters that include an element that we have classified above as a chromatin-altering element. Based on each of these promoters, we define a sequence context, i.e., 163 bp sequence from the native promoter in which we mutated annotated regulatory elements. We further define, for five of these promoters, a regulatory architecture, i.e., a combination of regulatory elements placed in the order and relative distance displayed by the native promoter (we employ consensus sites instead of the native sites). Even though promoter variants with these natively derived sequence contexts initially display different chromatin landscapes (mean pairwise correlation in occupancy patterns between contexts of \( r = 0.66 \)), this diversity is suppressed by embedding the different architectures (\( r = 0.77-0.93 \), Figure S6C). Embedding any of the examined architectures results, for example, in the eviction of an NFR nucleosome (Figure S6C). Correspondingly, expression is modulated more by the different architectures than by the different contexts, and the ranking of architectures by their expression is same in all of the examined contexts (Figure S6C).

Notably, this consistent ability of one architecture to drive higher expression than another might rely both on the composing elements’ identity and their arrangement. As our above results suggest, a specific ordering of the composing elements can be more beneficial in terms of nucleosome remodeling than another ordering. It is therefore possible that the importance of the elements’ order might depend on the context and the elements identity. Indeed, we find that when we permute the order of elements from the RPL28-dervied architecture, variants show similar expression ranking when elements are embedded in the open versus closed synthetic context (Spearman correlation...
By carrying out accurate binding and expression measurements on a large-scale, systematically designed library of promoter variants, we demonstrate how the readout of regulatory sequences relies on TF-nucleosome interplay. We reveal pronounced differences in the sensitivity of TFs to the chromatin context in which their motifs are embedded and classify TFs by their capacity to alter chromatin, promote expression, or perform both. As eukaryotic regulatory sequences are riddled with clusters of TFBSs, we then characterize how different combinations of sites give rise to different modes of nucleosome-mediated TF collaborations. We analyze homotypic clustering and further focus on heterotypic clustering of sites for TFs differing in their capacities. We specifically show how the expression outcome quantitatively depends on the different determinants of these collaborations, namely, the initial chromatin context, the conferred effect of one TF on chromatin, and the transcriptional capacity of the second benefiting TF. These inter-dependencies are reflected also in an analysis of promoter occupancy versus expression on the entire library. We find a positive correlation between expression and a more downstream position or lower occupancy of the +1 nucleosome (Figures S7F, S7I, S7J, and S7M) and to a lesser extent, with a more upstream position or a lower occupancy of the −1 nucleosome (Figures S7E, S7G, S7J, and S7K). Yet, we find that these chromatin features do not guarantee high expression. Similarly, while highly expressing variants show low occupancy of an NFR-nucleosome, the absence of such a nucleosome is not a guaranty for high expression (Figures S7H, S7J, and S7L), likely because expression further depends on the TFs that are now able to occupy the NFR.

Taken together, our results demonstrate the importance of going beyond the mere enumeration of composing TFBSs and accounting for ways in which they act in concert within different chromatin contexts. Our characterization thereby provides insights as to mechanisms that might underlie seemingly puzzling observations accumulated over recent years. These include observations, from yeast to higher eukaryotes, of TF binding events that do not contribute to expression, observations of differential binding and expression for similar TFBSs found in different genomic contexts, or observed constraints that are rarely mechanically understood on specific combinations and precise organization of TFBSs within regulatory sequences (Levo and Segal, 2014). Accounting and characterizing the quantitative contribution of TF-nucleosome interplay to gene expression further extends our synthetic biology toolbox, i.e., offering new strategies to manipulate regulatory sequences in order to obtain a desired outcome.
Figure 7. Dissecting the “Logic” of the CYC1 Promoter

(A) An illustration of the native CYC1 promoter variant is shown at the top. The occupancy pattern along this variant is shown below. Reads mapped to this variant were clustered (see the STAR Methods) to eight clusters. Shown is the mean occupancy pattern in each cluster, with the size of each row reflecting the relative proportion of this cluster. The clusters with three co-occurring nucleosome constitute 53% percent of the population.

(B) Each row shows the occupancy patterns and expression measurements for variants derived from the native CYC1 promoter; a few CpGs were added to a 163-bp sequence from the native promoter in order to improve the binding assay resolution (row 4). Illustrations on the left show the additional promoter manipulations carried out; replacement of the native HAP1 site by the SwissRegulon consensus site (Pachkov et al., 2007) (row 3) or by the reverse-complement of this consensus site (top row), improvement of the Hap2–Hap5 site with the UP1 mutation (Forsburg and Guarente, 1988), replacements of the Hap2–Hap5 site by sites for Rap1 or Abf1, mutations to the Hap1 or Hap2–Hap5 sites or to other regions previously suggested to contribute to expression (Forsburg and Guarente, 1998; Morohashi et al., 2008; Zhang and Guarente, 1994) (for sequences used, see Table S1).

(C) Each row shows the occupancy patterns and expression measurements for variants with combinations of the consensus sites for Hap1 and Hap2–Hap5. Sites are embedded either in the open chromatin context (left column) or the closed chromatin context (right column).
While the presented application of our assay is carried out in yeast, the principles emerging pertain to universal properties of regulatory sequences and regulatory proteins (e.g., TFBS clustering, TF-nucleosome competition for DNA accessibility, and TF-mediated nucleosome remodeling). Future applications can further examine these issues in higher eukaryotes, as our assay can be readily adjusted to diverse organisms. Such application can be facilitated by site-specific genomic integration of large-scale sequence libraries (Dickel et al., 2014). Nuclei extraction can replace the use of yeast spheroplasts (possibly improving the ability to capture TF footprints as it is less disruptive). Furthermore, the use of a GpC methyltransferase provides a means to not only circumvent the issue of endogenous methylation but rather simultaneously monitor both DNA methylation and occupancy (Kelly et al., 2012; Nabilsi et al., 2014). The assay can then be applied both to libraries of genomic sequences (Arnold et al., 2013; Dickel et al., 2014; Kheradpour et al., 2013; Murtha et al., 2014) or libraries of systematically designed variants (Fiore and Cohen, 2016; Smith et al., 2013). These can further focus on TFs that were previously suggested to collaborate (including tissue-specific factors [Smith et al., 2013] and pluripotency factors [Fiore and Cohen, 2016]) or facilitate chromatin remodeling (as pioneer factors [Zaret and Mango, 2016]).

Notably, our observations were afforded by the high-resolution and quantitative nature of our binding measurements allowing us, for example, to observe subtle changes in NFR width and in the fraction of the population in which a region is occupied. The resolution of our assay depends on the frequency of the motif recognized by the methyltransferase. Designed or genomically derived sequences altered to include this motif, or the use of multiple methyltransferases, thereby provide means to increase sensitivity. Furthermore, as the assay does not involve antibodies or size selection, it allows capturing binding events of differential size of both TFs and nucleosomes. Indeed, in our current application, we note that the observed NFR-nucleosomes, as we refer to them, often occupy a shorter region than the classical length of a nucleosome (147 bp), and further examination might find these related to recently characterized subnucleosomes (Henikoff et al., 2011), fragile nucleosomes (Kubik et al., 2015; Weiner et al., 2010; Xi et al., 2011), or pre-nucleosomes (Khuong et al., 2015).

While we focused here mostly on population level occupancy patterns, we also demonstrated how our assay facilitates single-cell analyses as it extracts distinct configurations of co-occurring binding events and estimates their frequency in the cell population. As is the case with the emergence of single-cell genomic approaches for measuring nucleosome occupancy (Buenrostro et al., 2015; Cusanovich et al., 2015), the degree of variability in binding between cells can be assessed and its determinants can be explored. We have already observed a large diversity in the presence and location of the three promoter nucleosomes occupying the same promoter variant in different cells. We find that a strong site for Abf1 or Gal4, for example, suppresses this variability as many cells now show a pronounced TF footprinting or an NFR. Future applications of this assay can be geared to such single-cell analyses. Applications focused on TFs that are more amendable to footprinting (e.g., as demonstrated here for Abf1 and Hap1) can further extend our understanding on the frequency of co-occurring TF binding events. Additional applications can be performed on subpopulations of cells differing in their expression (e.g., obtained by cell sorting), as opposed to the current study in which binding and expression measurements where performed separately on the entire population of cells carrying the library. This can allow a closer examination of the occupancy pattern in cells with the same promoter showing differential expression. With a design tailored for such purposes, these single-cell-oriented applications have the potential to allow us to extract the expression contribution of different binding configurations and relate variability in binding to variability in gene expression.

STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

+ KEY RESOURCES TABLE
+ CONTACT FOR REAGENT AND RESOURCE SHARING
+ EXPERIMENTAL MODEL AND SUBJECT DETAILS
+ METHOD DETAILS
  - Library Design
  - Library Preparation and Genomic Integration
  - Expression Measurements
  - Binding Measurements
+ QUANTIFICATION AND STATISTICAL ANALYSIS
  - Expression Measurements
  - Binding Measurements
+ DATA AND SOFTWARE AVAILABILITY
+ ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information includes seven figures and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2017.01.007.

AUTHOR CONTRIBUTIONS

M.L. conceived the study, designed the library, developed the experimental protocol, performed experiments and analysis, and wrote the manuscript. T.A.S. participated in the development of the binding assay and performed binding measurements. M.L.-P. constructed the strains and performed expression experiments. Y.K. participated in the development of the binding assay. A.W. assisted in the development of the experimental protocol. Z.Y. assisted in the design of the library and provided comments. E.S. conceived the project and experiments, supervised the work, and wrote the manuscript.

ACKNOWLEDGMENTS

This work was supported by grants from the European Research Council (ERC) and the NIH to E.S. M.L. thanks the Azrieli Foundation for the award of an Azrieli Fellowship. We thank Lucas Cary, Eilon Sharon, and the late Jonathan Wol-dom for their invaluable help in conceiving this project. We thank Einat Zalck-var, Leeat Keren, Shira Weingarten-Gabbay, and Eran Kotler for their insightful input. We thank Thomas Gregor and Mike Levine for their support in the completion of this project.


## STAR METHODS

### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chemicals, Peptides, and Recombinant Proteins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yeast Nitrogen Base without Amino acids</td>
<td>BD</td>
<td>Cat#291940</td>
</tr>
<tr>
<td>Histidine</td>
<td>Sigma</td>
<td>Cat#H6034-25G</td>
</tr>
<tr>
<td>L-Leucine</td>
<td>Sigma</td>
<td>Cat#L8912-25GR</td>
</tr>
<tr>
<td>Polyethylene glycol (50% PEG 3350)</td>
<td>Sigma</td>
<td>Cat#P3640-1KG</td>
</tr>
<tr>
<td>Lithium acetate</td>
<td>Sigma</td>
<td>Cat#517992-100G</td>
</tr>
<tr>
<td>Salmon sperm</td>
<td>Sigma</td>
<td>Cat#D9156</td>
</tr>
<tr>
<td>2% (final) from 37% Formaldehyde solution</td>
<td>J.T. Baker</td>
<td>Cat#7040</td>
</tr>
<tr>
<td>Glycine</td>
<td>Sigma</td>
<td>Cat#G7126-1KG</td>
</tr>
<tr>
<td>Phenylmethanesulfonyl fluoride ≥ 98.5% (GC) (PMSF)</td>
<td>Sigma</td>
<td>Cat#P7626-250MG</td>
</tr>
<tr>
<td>Benzylacetone (BZA)</td>
<td>Sigma</td>
<td>Cat#B3015</td>
</tr>
<tr>
<td>D-sorbitol, minimum 98%</td>
<td>Sigma</td>
<td>Cat#S1876-1KG</td>
</tr>
<tr>
<td>β-mercaptoethanol</td>
<td>Sigma</td>
<td>Cat#M6250-100ML</td>
</tr>
<tr>
<td>Lyticase from Arthrobacter</td>
<td>Sigma</td>
<td>Cat#L2524-200KU</td>
</tr>
<tr>
<td>Nonidet P–40 (IGEPAL CA-630)</td>
<td>Sigma</td>
<td>Cat# 21-3277 SAJ</td>
</tr>
<tr>
<td>15% Ficoll</td>
<td>Sigma</td>
<td>Cat# F2637-5G</td>
</tr>
<tr>
<td>Gel Loading Dye Orange (6x)</td>
<td>NEB</td>
<td>Cat#B7022S</td>
</tr>
<tr>
<td>GelStar Nucleic Acid Gel Stain 10,000x concentrate</td>
<td>LONZA</td>
<td>Cat#50535</td>
</tr>
<tr>
<td>2-Propanol AR-b</td>
<td>Bio-Lab ltd.</td>
<td>Cat#001626052100</td>
</tr>
<tr>
<td>Sodium acetate, anhydrous</td>
<td>Sigma</td>
<td>Cat#S8750-500G</td>
</tr>
<tr>
<td>Glycogen, RNA grade 20mg/ml, 0.1ml</td>
<td>Fermentas</td>
<td>Cat#R0551</td>
</tr>
<tr>
<td>Tris, Ultra Pure</td>
<td>MP</td>
<td>Cat#819620</td>
</tr>
<tr>
<td>EDTA, Disodium Salt, Dihydrate, Crystal</td>
<td>J.T. Baker</td>
<td>Cat# 8993-01, 500G</td>
</tr>
<tr>
<td>Sodium dodecyl sulfate 20%</td>
<td>Sigma</td>
<td>Cat# 001981232300</td>
</tr>
<tr>
<td>Proteinase K from Tritirachium album</td>
<td>Sigma</td>
<td>Cat#P6556</td>
</tr>
<tr>
<td><strong>Critical Commercial Assays</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XmaJI (AvrII) restriction enzyme</td>
<td>Thermo Scientific</td>
<td>Cat# ER1561</td>
</tr>
<tr>
<td>SexAI restriction enzyme</td>
<td>Thermo Scientific</td>
<td>Cat#FD2114</td>
</tr>
<tr>
<td>Cpol (RsrII) restriction enzyme</td>
<td>Thermo Scientific</td>
<td>Cat# ER0741</td>
</tr>
<tr>
<td>SgsI (AscI) restriction enzyme</td>
<td>Thermo Scientific</td>
<td>Cat# ER1891</td>
</tr>
<tr>
<td>YeaStar Genomic DNA Kit</td>
<td>ZYMO RESEARCH</td>
<td>Cat#D2002</td>
</tr>
<tr>
<td>Herculase II Fusion DNA polymerase</td>
<td>Agilent Technologies</td>
<td>Cat#600675</td>
</tr>
<tr>
<td>ZR-96 DNA Clean &amp; ConcentratorTM-5</td>
<td>ZYMO RESEARCH</td>
<td>Cat#D4024</td>
</tr>
<tr>
<td>electrophoresis Midi GeBAflex tube</td>
<td>Gene Bio-Application L.T.D</td>
<td>Cat#D010-100</td>
</tr>
<tr>
<td>High Sensitivity D1000 ScreenTape</td>
<td>Agilent Technologies</td>
<td>Cat#5067-5584</td>
</tr>
<tr>
<td>cComplete Mini EDTA free</td>
<td>Roche</td>
<td>Cat#1183617001</td>
</tr>
<tr>
<td>MSsII</td>
<td>Thermo Scientific</td>
<td>Cat# EM0821</td>
</tr>
<tr>
<td>RNase</td>
<td>Sigma</td>
<td>Cat# R5500-100MG</td>
</tr>
<tr>
<td>EZ DNA methylation kit</td>
<td>ZYMO RESEARCH</td>
<td>Cat# D5005</td>
</tr>
<tr>
<td>KAPA HiFi HotStart Uracil+ ReadyMix PCR (2x)Kit</td>
<td>KAPA Biosystems</td>
<td>Cat#KM2801</td>
</tr>
<tr>
<td>QIAquick PCR Purification Kit</td>
<td>QIAGEN</td>
<td>Cat#28104</td>
</tr>
<tr>
<td>MiSeq Reagent Kit v3</td>
<td>Illumina</td>
<td>Cat#MS-102-3001</td>
</tr>
<tr>
<td><strong>Deposited Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw sequencing reads were deposited</td>
<td>NCBI Gene Expression Omnibus</td>
<td>GEO: GSE92300</td>
</tr>
</tbody>
</table>
CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents may be directed to, and will be fulfilled by the corresponding author Eran Segal (eran.segal@weizmann.ac.il).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All strains used in this study were derivatives of S. cerevisiae Y8205 (Tong and Boone, 2006), kindly provided by Charles Boone. Construction of strains with synthetic promoters driving YFP reporter are described below.

METHOD DETAILS

Library Design

The designed library consist of ~1600 sequences of length 210bp. Each sequence includes 19bp 5’ primer sequence, a 10bp unique barcode sequence, 163bp variable region and 18bp 3’ primer.

- 5’ primer sequence: 5’-GGGACCAGGTGCCGTAACG-3’
- 3’ primer sequence: 5’-TGATCGCCCTAGGATCGC-3’

5’ and 3’ primers include site for restriction enzymes SexAI and AvrII, correspondingly (underlined in the above sequences).

Barcodes were selected randomly, though we ensure no occurrence of CpGs, or stretches of more than 4 consecutive occurrence of the same nucleotide. Final barcodes differ from one another by at least 2bp.

Variable Region Description

The designed library includes two sections: the native section (consisting of ~600 sequence variants) and the synthetic section (consisting of ~900 sequences). The native section includes 9 subsets, each with variants derived from a different yeast promoter (ENO1, PGK1, HIS3, HIS4, RPL28, CYC1, GAL1, RPL4A, RPS28A, in the manuscript we focus mainly on the first 6 promoters). 163bp from the original promoter, thought to contain important regulatory elements based on previous literature, were selected (and are included as a variant in the library). If needed a few CpGs were carefully added to the sequence to improve the binding assay resolution, while attempting to avoid disruptive changes – the resulting sequences are referred as the context sequence for this set. Variants within

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experimental Models: Organisms/Strains</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. cerevisiae: Strain background: Y8205</td>
<td>(Tong and Boone, 2006)</td>
<td></td>
</tr>
<tr>
<td>Yeast master strain</td>
<td>(Raveh-Sadka et al., 2012)</td>
<td></td>
</tr>
<tr>
<td>Recombinant DNA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AgilentLib Hiscore plasmid</td>
<td>(Sharon et al., 2012)</td>
<td></td>
</tr>
</tbody>
</table>

| **Library amplification primer:** | | |
| The 5’ primer: 5’-GGGACCAGGTGCCGTAACG-3’ | | |
| The 3’ primer: 5’-TGATCGCCCTAGGATCGC-3’ | | |

| **Amplification of sequences from expression bins:** | | |
| The 5’ primer: 5’-XXXXXGGGACCAGGTGCCGTAAC-3’ (where the Xs represent the bin’s unique sequence) | | |
| The 3’ primer: 5’-NNNNNTTATGTGATAATGCCTAGGATCGC-3’ (where the Ns represent random nucleotides) | | |

| **Post bisulfite conversion PCR 5’ primer,** | | |
| 5’-XXXXXATAAAGAAATAGACCTAGATAGGGTT-3’ (where the Xs represent a sample barcode) | | |
| 5’-NNNNNCATCACCATCTAATTTAACCATAAAAT-3’ | | |

| **Data analysis done using MATLAB** | | |
| Mathworks | | |
each of the native subsets mostly include careful manipulations to previously annotated regulatory elements (see Table S1); mutations to binding site content (e.g., employing mutations that were shown to significantly reduce or abolish TF binding in vitro, or reduce the expression of a regulated gene in vivo), binding site orientation, location, and replacement by other sites.

The synthetic section of the library (consisting of ~900 sequence variants) provides means for a more systematic and controlled examination of TF activity in different chromatin contexts. In these variants we mostly place various combinations of consensus binding elements derived from the HIS3 yeast promoter and a closed chromatin context derived from the GAL1-10 promoter (after the removal of characterized regulatory elements and the careful addition of a few CpGs).

Each context variant (i.e., 9 for the native sets and 2 for the synthetic sets), appears in the design 10 times, with a different 10bp barcode (designed not to include CpGs and found ~285 to ~276bp upstream of the TSS after genomic integration). These differential barcode sets serve as internal controls, facilitating the assessment of experimental reproducibility of both binding and expression measurements (Figure S1).

**Library Preparation and Genomic Integration**

The designed sequences were synthesized on Agilent programmable microarrays (LeProust et al., 2010), and cloned into the pKT103-based plasmid as described elsewhere (Sharon et al., 2012), except for the following changes: For library amplification 3 ng of DNA was used instead of 12 ng and, 8 cycles of PCR were performed (see primer sequences in the above library design section). 250 ng was then used for the digestion of library with XmaI (AvrII, Catalog No. ER1561, Fermentas) and SexAI (CsiI, catalog No. FD2114, Fermentas) restriction enzymes; for 2.5 hr at 37°C followed by inactivation of the enzymes for 5 min at 65°C. Reaction mixture contained 8 μL FD buffer, 2 μL of each enzyme, 4 μL alkaline phosphatase (Fermentas) and 4.6 μg of the plasmid in a total volume of 80 μL.

To obtain genomic integration, we then digested the plasmids with the designed library, and transformed them into our yeast master strain (described elsewhere (Raveh-Sadka et al., 2012)). Digestion was performed with Cpol (RsrII, Catalog No. ER0741) and SgsI (Ascl, Catalog No. ER1891) (Termo Fisher Scientific, Fermentas) restriction enzymes, in large scale – 192 reactions in two 96-well plates. The digestion–reaction mixture contained per each reaction: 2 μL Fast Digest buffer (supplied by Fermentas), 1 μg of the library-plasmid, 0.5 μL of each enzyme, 1 μL of FastAP (#EF0654, Fermentas) and double distill water (DDW) up to a total volume of 20 μL. The mixture was incubated for 2.5 hr at 37°C, followed by 20 min inactivation at 65°C.

A day before transformation, the culture of the master strain was grown in 1200 mL of YPD medium in three Erlenmeyer’s of 400 ml. At the day of transformation, cells reached an optical density of 0.9 (OD600), 10⁸ cells per one OD unit, and were split into four buckets of 300 mL each. Centrifugation at 4°C was performed at 4000 rpm for 5 min and the supernatant was discarded. Cells were washed with 300 mL of DDW and split into two 50 mL tubes. Each pellet of cells were resuspended at 4°C with 40 mL of DDW and centrifuged at 3000 rpm for 5 min and the supernatant was discarded.

The transformation mixture was prepared in advance for 216 reactions and each reaction contained: 120 μL of 50% PEG 3350 (Sigma), 18 μL of 1M LiAC (Catalog No. 517992, Sigma Aldrich), 5 μL of 10 mg/ml Dextranomucic acid from salmon testes (Catalog No. D9156, Sigma Aldrich), 20 μL digestion library-plasmid and DDW up to 180 μL.

All steps from this point were carried on out using a programmable robotic system (Tecan Freedom Evo, Tecan Inc.). Cells were resuspended in the transformation mixture and distributed into two 96-well plates, 160 μL in each well; then, 20 μL of digested library-plasmid DNA was added to each well. The cells and the DNA were shaken to make a homogeneous mixture and then incubated for 30 min at 30°C. After incubation, the plates were shaken again and incubated for 40 min at 42°C in a water bath.

The two 96 plates were then centrifuged at RT, at 4000 rpm for 1.5 min and transformation-mixtures was discarded. Cells were washed with 160 μL DDW per well; then vortexed and centrifuged at 4000 rpm for 1.5 min, to obtain pellets. Resuspension of each well was performed with 100 μL DDW, followed by shaking. The eluted cells were collected into one pool and split into three Erlenmeyers containing 93 mL of SCD-URA (synthetic complete media with 2% glucose and without uracil), 7 mL of cells per Erlenmeyer. Transformed cells (120,000 transformants) were grown at 30°C for 72 hr until the culture reached the stationary phase.

After genomic integration the loci of interest includes the following 520bp:

```
TAAAGAATAGACTAGATGGTTTGCCTGTCCCTCCGTTTGGAAACAGGGTTCCACTATTCAGAAGGACTCTACGTTAAAGACGTAATCGGAGTTTTTTTTGGGACCAGGTGGCGTAACG
```

[10bp barcode + 163bp variable region]

```
GCGATCTCAGGATCCATTACATGATTACGTATATCAGTATATAAACGTAAATCTCGATTGTTTCCTTGAGAATATACTAAACATGAGCAAGGAAAGGATTGTCATAGGCGAAGATTATATTCCAGG
```

**Expression Measurements**

Expression measurements were carried similarly to what was previously described (Sharon et al., 2012). In brief, to adapt cells to the medium for sorting, cells were grown to stationary phase in in SC-Gal-URA (synthetic complete media with 2% galactose and without
uracil) medium without amino acids, except for histidine and leucine. The culture was diluted again in similar medium (to OD<sub>600</sub> 0.03–0.05) and grown to the mid-exponential phase (OD<sub>600</sub> 0.5–1.5) for sorting. Sorting was performed with the FACSAria cell sorter (Becton Dickinson). We first gated cells based mCherry fluorescence, to ensure and a single integration event of our cassette (and to obtain a relatively homogeneous size population). We then sorted cells to 16 bins based on YFP fluorescence. In each sorting strategy, the expression bins contained equal fractions of the library cells, and we collected a total of over 10,000,000 cells. Sorted cells were grown in SCD-URA medium to stationary phase. We then performed DNA extraction of each bin, using YeaStar Genomic DNA Kit, (ZYMO RESEARCH, THE Epigenetics COMPANY, Catalog No: D2002), followed by PCR amplification. The reaction mixture contained 6 μL 5×Herculase II reaction buffer, 6 μL 2 mM dNTPs mix, 30 ng DNA, 1 μL Herculase II Fusion DNA Polymerase (Agilent), 3 μL 10 μM 3’ and 5’ primers and DDW up to 30 μL. PCR was carried out as follows; 95 °C for 5 min, 24 cycles of 95 °C for 20 s, 72 °C for 1 min and one cycle of 72 °C for 5 min. The 3’ primer was common to all bins (5’-NNNNNTTATGTGATAATGCCTAGGATCGC-3’, where the Ns represent random nucleotides). The 5’ primer had a common sequence and a unique upstream 5-bp barcode sequence (underlined) that was specific to each bin (5’-XXXXXXXXXXXXXXXXX-3’, where the Xs represent the bin’s unique sequence). Three 5’ primers were used for the amplification of each bin, reads were later joined after we found them to be highly consistent. Amplified product was purified by ZR-96 DNA clean & concentrator – 5 (ZYMO RESEARCH, THE Epigenetics COMPANY, Catalog No: D4024), and run on gel; Fragments of the correct size were cut from the gel, eluted using electroelution Midi GeBAflex tubes, and precipitated using the standard sodium acetate and isopropanol protocol. Ten nanograms were used in library preparation for sequencing (protocol adopted from Blecher-Gonen et al., 2013). The DNA was amplified using 14 amplification cycles and library was analyzed using the 2200 TapeStation system and sequenced with 100-bp, paired-end reads flowcell on an Illumina HiSeq 2000 sequencer.

**Binding Measurements**

**Experimental Procedure**

**Spheroplasts Preparation and M.SsSI Treatment.** Strains containing different promoter variants were cultivated at 30 °C in 300 mL of SC-Gal (synthetic complete media with 2% galactose) medium without amino acids, except for histidine and leucine. Exponentially growing cells (OD<sub>600</sub> ~0.4) were harvested, and cross linked by adding 2% Formaldehyde for 5 min on shaker. Notably extended crosslinking times, as more common in ChIP assays can improve the ability to capture TF binding events, yet occupancy patterns might reflect time-averaging. Formaldehyde was quenched by shaking cells with 1/10 volume 1.25M freshly made Glycin for additional 5 min. Fixed cells were centrifuged (4,000 RPM for 6 min, 4 degrees) and washed with double-distilled water twice. The pellet then re-suspended in 1ml of spheroplasting buffer and incubated for 40 min, shaking at 30 °C. 1ml of spheroplasting buffer contained 849.34ul 1M sorbitol, cOmplete Mini EDTA free (Roche, 11836170001) dissolved in Sorbitol 1mM (1 tablet was dissolved in 1.5 mL for 10 mL of spheroplasting buffer), 0.66ul of 1.43M β-mercaptoethanol and 2.48mg/ml mg of lyticase (Sigma, L2524-200KU, amount used is equivalent to 6000U). We tested that spheroplasts were obtained by OD reduction in water using 10 μl of the sample. Spheroplasts were then centrifuged (3,700 RPM for 5 min) and washed twice with 1 M PI-sorbitol. The pellet was resuspended with 142 μl of CpG Methyltransferase M.SsSI buffer (1X M.SsSI buffer, 1X SAM, 2M sorbitol, cOmplete Mini EDTA free and 0.375% Nonidet P–40 (igepal; Sigma, CA-630)), incubated for 10 min at 30 °C and then treated with M.SsSI by addition of 8U (Thermo Fisher Scientific, EM0821) for 15 min at 30 °C. Digestions were terminated with 150 μl stop solution (same volume as the sample, 10 mM EDTA and 1% SDS). Samples were incubated with 0.18 mg/ml RNase (Sigma, R5500-100MG) that was added to the stop solution for 1 hr at 37 °C. To reverse cross-link 5 μl (0.1 mg) of Proteinase K (Sigma, P6556, 20mg/ml) were added to the samples and incubate over-night at 65 degrees (in dry block). DNA was extracted using Phenol/Chloroform purification (2 rounds of chloroform) and eluted in 30 μl DDW.

**Bisulfite Conversion**

Bisulfite treatment was carried out using the EZ DNA methylation kit (Zymo research, D5005): to 1 μg of DNA in 45 μl DDW, 5 μl of M-Dilution buffer were added. Sample (50 μl) was mixed and incubated for 15 min at 37 °C. Then 100 μl CT Conversion Reagent was added to each sample (CT Conversion Reagent was prepared by adding 750 μl water, 210 μl M-Dilution buffer, vortex and rotating for 10 min at room temp) and the mix was rotated for 10 min at room-temp. Each sample (150 μl) was divided to 3 PCR tubes and incubated in a thermocycler using the following protocol: 20 cycles of [30 s at 95 °C, 15 min at 50 °C] and re-united to one tube when incubation was done. To each sample 400 μl of M-Binding buffer was added and mixed, then loaded on a Zymo-Spin IC Column. Zymo-Spin IC Column was centrifuged briefly (30 s) and washed with 200 μl M-Wash buffer (prepared according to the manufacturer’s instructions). Next, 200 μl of M-Desuphonation buffer was added on column, incubated for 20 min at room temp, and washed again with 200 μl M-Wash buffer. Second wash was performed using 150 μl M-Wash buffer. Finally, sample was eluted in 12 μl water.

**Post-conversion PCR**

Two PCR reactions (each with 5.5 μl of the sample and 0.75 μl of each primer at 10 μM, in a total volume of 25 μl) were performed. Primers used (matching regions are underlined in the above sequence, obtained after genomic integration)

- 5’ primer – 5’-XXXXXXXXTAAAAAAAAATAGACCTAGATAGGGTT-3’ (where the Xs represent a sample barcode, if several samples were multiplexed)
- 3’ primer – 5’-NNNNNCATCACCATCTAATTCACAAAAT-3’
PCRs were carried out using KAPA HiFi HotStart Uracl+ ReadyMix PCR Kit (KAPA Biosystems, KK2801). PCR conditions were as follows: initial denaturation at 95°C for 3 min., denaturation at 98°C for 20 s, annealing at 50°C for 30 s and extension at 72°C for 60 s for 30 cycles. PCR products were purified using QIAquick PCR Purification Kit (QIAGEN, 28104) and eluted in 30 μl water. Size of product was verified using the 2200 TapeStation system (Agilent, 5067-5584).

To separate desired band, 200 ng of each sample was mixed with 15% Ficoll (Sigma)/Gel Loading Dye Orange (NEB-B7022S) and loaded on 2% agarose gel stained with GelStar (Lonza). The bands were cut under UVIblue blue light transilluminator (UVITEC). The DNA was eluted from the gel using electroelution Midi GeBAflex tubes (Gene Bio-Application), precipitated with 1 volume isopropanol, 1/10 volume 3 M NaOAc (pH 5.2), and 0.1 μg/μL glycogen (Fermentas) overnight at −20°C, and resuspended in 10 μl Tris 10 mM pH = 8.

**High-Throughput Sequencing**

Ten nanograms were used in library preparation for sequencing (protocol adopted from Blecher-Gonen et al., 2013). The DNA was amplified using 14 amplification cycles and library was analyzed using the 2200 TapeStation system. Libraries were sequenced on a 300 paired-end flow cells (MiSeq Reagent Kit v3, MS-102-3001) on MiSeq desktop sequencer (Illumina).

**QUANTIFICATION AND STATISTICAL ANALYSIS**

**Expression Measurements**

Mean expression per sequence variant was obtained based on a weighted average of the variant’s reads in each expression bin, as previously described (Sharon et al., 2012).

**Binding Measurements**

**Mapping and Initial Analysis**

Sequence reads were mapped to library variants initially based on the 10bp barcode included in the design (this is performed for reads that were found to include the 5’ primer sequence after bisulfite conversion). We then compute an alignment of each read to the reference (i.e., designed) variant, assuming full conversion (that is every C in the reference sequence is expected to appear as a T in the corresponding read), and disregarding the status of Cs within CpGs. Reads with more than 1.5% sequencing errors (including mutations, gaps or Ns) are discarded. Reads with less than 95% conversion efficiently are discarded.

We then call a methylation status for each CpG; methylated if we find a C in that position, unmethylated if we find T, or NaN if there is uncertainty (N in the sequencing or disagreement between the paired-end reads). Notably a large fraction of the variability designed region (120bp out 163bp) is covered by reads from both ends and we use this information to improve the quality of our alignment and methylation status calling. We then compute for each CpG position in each variant the fraction of reads in which it was unmethylated, and use this as a proxy of occupancy in that position. To obtain full occupancy pattern along the promoter we interpolate this data. The resulting vector, encompassing values for 456bp, from the first occurrence of a CpG to the last occurrence, is the one presented and use this as a proxy of occupancy in that position. To obtain full occupancy pattern along the promoter we interpolate this data. The resulting vector, encompassing values for 456bp, from the first occurrence of a CpG to the last occurrence, is the one presented and use this as a proxy of occupancy in that position. To obtain full occupancy pattern along the promoter we interpolate this data.
DATA AND SOFTWARE AVAILABILITY

The accession number for the raw data reported in this paper is NCBI GEO: GSE92300. Processed data, i.e., occupancy patterns and mean expression per variant can be found in Table S2.

ADDITIONAL RESOURCES

A few additional examples for systematic manipulation of regulatory elements in variants derived from native yeast promoters can be found in: https://genie.weizmann.ac.il/transcription_factor_activity.html.