# Core promoter sequence in yeast is a major determinant of expression level

Shai Lubliner[1,#], Ifat Regev[1,2,#], Maya Lotan-Pompan[1,2], Sarit Edelheit[3], Adina Weinberger[1,2,*], Eran Segal[1,2,*]

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

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

[3]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel.

[#]These authors contributed equally to this work

[*]Corresponding authors: eran.segal@weizmann.ac.il, adina.weinberger@weizmann.ac.il

## Abstract

The core promoter is the regulatory sequence to which RNA polymerase is recruited and where it acts to initiate transcription. Here, we present the first comprehensive study of yeast core promoters, providing massively parallel measurements of core promoter activity and of TSS locations and relative usage for thousands of native and designed sequences. We found core promoter activity to be highly correlated to the activity of the entire promoter, and that sequence variation in different core promoter regions substantially tunes its activity in a predictable way. We also show that location, orientation and flanking bases critically affect TATA element function, that transcription initiation in highly active core promoters is focused within a narrow region, that poly(dA:dT) orientation has functional consequence at the 3' end of promoters, and that orthologous core promoters across yeast species have conserved activities. Our results demonstrate the importance of core promoters in the quantitative study of gene regulation.

## Introduction

The RNA polymerase II (Pol II) core promoter is the region to which Pol II and its accompanying general transcription factors are recruited to the DNA, form the pre-initiation complex (PIC) and act to initiate transcription (Smale and Kadonaga 2003). In yeast, the PIC is recruited to a TATA element, either a consensus TATA box or a weaker one with 1-2 mismatches to the consensus (Singer et al. 1990; Basehoar et al. 2004; Sugihara et al. 2011; Rhee and Pugh 2012). In both yeast and metazoans PIC recruitment to a TATA element leads to promoter DNA melting ~20 bp downstream of it (Giardina and Lis 1993), with the promoter sequence ~30 bp downstream of the TATA element being at the Pol II active center (Bushnell et al. 2004; Miller and Hahn 2006). While in metazoans this leads to transcription initiation ~30 bp downstream of the TATA element, in *S. cerevisiae* Pol II performs a downstream scan of the template strand in search of transcription start sites (TSSs) (Giardina and Lis 1993; Kuehner and Brow 2006; Sugihara et al. 2011; Fishburn and Hahn 2012) in a manner that depends on the sequence around and upstream of the TSS (Hahn et al. 1985; Furter-Graves and Hall 1990; Faitar et al. 2001; Zhang and Dietrich 2005; Fishburn and Hahn 2012; Goel et al. 2012). This results in transcription initiation between 40-120 bp downstream of the TATA element, and typical core promoter lengths of 100-200 bp (Smale and Kadonaga 2003; Lubliner et al. 2013). To allow access of the PIC to the DNA, core promoters typically contain a nucleosome free region (NFR) (Field et al. 2008; Kaplan et al. 2009).

In the study of regulatory sequences and their effects on expression, core promoter sequences remained relatively understudied, as most efforts are directed at transcription factor (TF) binding sites and their role in determining regulatory logic and expression levels (Levo and Segal 2014). In yeast, many core promoter related studies revolved around the effects of TATA sequence specificity on expression (Chen and Struhl 1988; Singer et al. 1990; Mahadevan and Struhl 1990; Stewart and Stargell 2001; Basehoar et al. 2004; Mogno et al. 2010; Rhee and Pugh 2012). One study explored how variation at the TSS region affected TSS efficiency and expression in a single promoter (Kuehner and Brow 2006). In a recent computational study we showed that various yeast core promoter features, such as T-content upstream and A-content at and downstream of the main TSS, correlate with maximal promoter activity (Lubliner et al. 2013), yet proving causality required experimental validation. To date, no experimental study

comprehensively explored the role of the core promoter region in determining overall promoter activity or the effects of core promoter sequence features on expression level and on TSS selection.

Here, we extended a powerful experimental system (Sharon et al. 2012) by designing a library of thousands of native and synthetic core promoter sequences (synthesized by Agilent Technologies) that drive *in vivo* expression of a reporter gene within *S. cerevisiae*, under the regulation of a constant upstream region coming from a strong constitutive promoter. Our results provide new insights into the role of core promoters in yeast and highlight their pivotal role in the regulation of transcription.

# Results

## Pooled measurement of expression and of transcription start sites for thousands of designed core promoter sequences

In order to explore the effects of core promoter sequence on expression level and on the distribution of alternative TSSs, we adapted a method previously developed in our lab (Sharon et al. 2012) (**Fig. 1** and Methods). We designed *in-silico* and then synthesized a library of 13,000 DNA oligos, each containing a 118 bp long core promoter sequence, driving expression of a yellow fluorescent protein (*YFP*). The library consisted of 7,536 unique core promoter variants, several hundreds of which native, and the rest synthetically manipulated to explore various hypotheses. For the purpose of TSS location measurements (see below), for 5,464 of the 7,536 unique variants we also designed a second oligo, additionally containing a unique barcode encoded by synonymous mutations within the first 36 bases of the *YFP* sequence. We will refer to these 5,464 oligos as YFP-barcoded, and to the other 7,536 oligos as non-YFP-barcoded.

All oligos were cloned into a pool of plasmids, and plasmids were transformed into *S. cerevisiae* cells. In each plasmid the 118 bp long core promoter region was upstream of an intact *YFP* sequence, and downstream of the [-528,-129] region (positions relative to the translation start site) of the *RPL28* promoter, a constitutively expressed ribosomal protein (RP) promoter that includes a tandem pair of binding sites for the Rap1 transcription factor, the main regulator of yeast RP promoters (Lieb et al. 2001), as well as binding sites for Fhl1 and Sfp1 which are also known to regulate RP gene expression level (Zeevi et al. 2014). This region of the *RPL28* promoter does not contain TATA elements downstream of the Rap1 sites, limiting its competition with the core promoter variant over PIC recruitment.

For filtering and normalization purposes, the plasmid also contains a strong promoter driving the expression of a red fluorescent protein (*mCherry*). Expression measurements were performed by fluorescence activated sorting (FACS) of the yeast cells into 16 bins of YFP/mCherry levels, next-generation sequencing of the variable core promoter region, mapping the non-YFP-barcoded sequencing reads (while discarding YFP-barcoded reads, to avoid the effect of the barcode mutations on expression) to their respective core promoter sequences and YFP/mCherry bins and then computation of the mean YFP/mCherry level of each of the 7,536 non-YFP-

barcoded core promoters from the mapped reads (**Fig. 1**). Since we measured protein levels, and were interested in transcriptional effects on expression, we included a constant A-rich 10-mer between the core promoter variant and the *YFP* sequence, since the bases immediately upstream of the translation start site were shown to greatly affect translation efficiency (Dvir et al. 2013). Further, all native core promoters included in our library were previously shown to have short 5' UTRs (see below) and most synthetic core promoters were based on native ones that were manipulated upstream of the 5' UTR. Thus, for the vast majority of sequences post transcriptional effects are expected to be negligible, with variation in expression measurements across the library representing transcriptional variation. The cause for this transcriptional variation was the sequence variation between the 118 bp long core promoter variants, and we therefore term our YFP/mCherry measurements 'core promoter activity'. The sequences and the measured activities of all 7,536 core promoter variants appear in **Supplementary Table 1**.

TSS measurements were performed by extracting total RNA from the yeast cells, 5' rapid amplification of cDNA ends (5' RACE) using primers specific to the *YFP* sequence, next-generation sequencing of the products, and then mapping the YFP-barcoded sequencing reads to their respective core promoter sequences, and computing TSS positions and relative abundances for each core promoter (**Fig. 1**). The TSS measurements of all 5,464 YFP-barcoded core promoters appear in **Supplementary Table 2**.

## Core promoters are major determinants of promoter activity

A major question is the extent to which core promoter sequence affects expression levels, since most yeast promoters are more than 500 bp long, with TF binding sites typically situated upstream of the core promoter. To address this question, we included in our library the [-118,-1] region (positions relative to the translation start site) of native promoters for which the activity of their entire promoter was reported, and for which previous studies reported their main TSS to be within 50 bp of the translation start site (Miura et al. 2006; Nagalakshmi et al. 2008).

One such set of 238 core promoters originates from orthologous ribosomal protein (RP) promoters of 4 *Saccharomyces sensu stricto* species (*S. cerevisiae*, *S. paradoxus*, *S. mikatae* and *S. bayanus*), representing 67 distinct RP genes, for which we recently measured the activity of

the entire promoter within *S. cerevisiae* (Zeevi et al. 2014) (see **Supplementary Table 3**). Notably, we found a very high correlation between our measurements of the activity of the core promoter and the activity reported for the entire promoter ($r$=0.78, $P<10^{-49}$, **Fig. 2A**). While most of the 238 core promoters come from promoters that are natively regulated by Rap1 (Tanay et al. 2005), the main upstream regulator in our experimental system, 15 of the 238 core promoters originate from 4 genes (*RPL3*, *RPL4A*, *RPL38*, *RPS28A*) not natively regulated by Rap1 (Lieb et al. 2001), yet they too showed the same trend ($r$=0.8, $P<10^{-3}$, **Fig. 2A**). Thus, the above high correlation is not due to any specific optimization of RP core promoters to interact with Rap1. Another set of 133 core promoters originates from *S. cerevisiae* promoters for which another study measured the activity of the entire promoter and classified them as constitutive (Keren et al. 2013) (see **Supplementary Table 4**). Here too, we found a high correlation ($r$=0.59, $P<10^{-13}$, **Fig. 2B**).

After exploring entire native promoters, we next sought to explore effects of varying different core promoter subsequences on expression. We manually annotated subsequences within native [-118,-1] promoter regions of 7 genes (*RPL3*, *RPL28*, *RPL25*, *RPL4B*, *GAL7*, *HSP12* and *RPB10*, see **Supplementary Table 5**), most of which are known to have a high level of promoter activity in at least one condition (Keren et al. 2013) (see illustration in **Fig. 3E**). In a 5' to 3' order, the annotated regions included the sequence upstream of a TATA element; a strong TATA element or a short region with weak TATA elements (Sugihara et al. 2011; Rhee and Pugh 2012); the 20-30 bp downstream of the TATA where the PIC is formed and initially unwinds the DNA (termed the PIC region) (Giardina and Lis 1993); a region through which Pol II is expected to pass while scanning for possible TSSs (termed the Scanning region, annotated for 3 out of 7 genes) (Lubliner et al. 2013); and the A-richer downstream region where the TSSs are located (termed the Initiation region) (Nagalakshmi et al. 2008; Lubliner et al. 2013). For each of the PIC, Scanning and Initiation regions we designed a set of mutated sequences, constructed by replacing the entire native region with equally long randomly generated sequences with several different base content biases. In addition, we produced a large set of sliding window mutations, where we mutated non-overlapping 10 or 30 bp long windows along various native [-118,-1] promoter regions included in our library. We measured the effect of these 4 sets of mutations on native core promoter activity and learned linear models predicting these effects based on sequence feature differences between mutated and native sequences. We

utilized a *K*-fold cross validation scheme (with *K*=5 or *K*=10) using *K* different partitions of the data into a training set and a held out test set, such that each mutant appeared once in a held out test set and *K*-1 times in a training set. For each of the *K* partitions of the data, a linear model was learned on the training set and its performance was assessed on the held out test set (see Methods). This approach enabled us to highlight sequence features that were consistently included into the different models as well as provide an estimate of how well these models explain the variation within the held out test sets.

Notably, for all 4 sets we found that mutations had a large and continuous spectrum of effects on expression (**Fig. 3A-D**). In addition, for all sets the learned linear models explained a large fraction of the variation in the test data, between 40% for the sliding window mutations and 72% for the Scanning region mutations (**Fig. 3A-D**). These results strongly suggest that the learned sequence features are causal and that tuning these features changes expression in a predictable way.

Most of the features found to predict core promoter activity (**Supplementary Figs. 1-4**) could be attributed to a few classes, illustrated in **Figure 3E**. These include not only signals that affect PIC recruitment (TATA elements and the core promoter's G\C content that affects nucleosome occupancy), but also signals expected to affect the efficiency of transcription initiation by Pol II, such as T/C-rich *k*-mers at the Scanning region and A-content at the Initiation region. See **Supplementary Note** for further discussion.

Taken together, the above results show that core promoter sequence is a major determinant of the activity of the entire promoter, and accordingly sequence variation within different core promoter regions substantially affects expression in a predictable way.

## Location, orientation, and flanking bases are important for TATA element function

Although the effect of variation in TATA element sequence on expression was previously studied (Chen and Struhl 1988; Singer et al. 1990; Mahadevan and Struhl 1990; Stewart and Stargell 2001; Basehoar et al. 2004; Mogno et al. 2010; Rhee and Pugh 2012), the effect of the location of the TATA element received little attention. To study this, we first examined the effect

of varying TATA box positions in a synthetic setting. As background sequences, we chose the native [-118,-1] regions of the *PDC1* and the *ENO2* promoters since they are highly expressed (Keren et al. 2013), have a TATA element further upstream and not within their [-118,-1] region, and their main TSS was reported to be 30 bp upstream of the ORF (Miura et al. 2006). Two different TATA consensus 8-mers (Basehoar et al. 2004) were used, TATAAAAA and the palindromic TATATATA. Each synthetic TATA insertion was generated by replacing an 8-mer on one of the background sequences with one of the TATA 8-mers. TATA insertion positions were denoted by the 8-mer's first position, and included all 50 positions in [-118,-69], and also positions -59, -49, -39, -29 and -19.

Similar results were obtained for both TATA 8-mers over both background sequences (see **Supplementary Fig. 5**). Below we focus on those for TATAAAAA insertions into the *PDC1* background (**Fig. 4A**). In line with the fact that TSSs are natively found at least ~40 bp downstream of the TATA element (Smale and Kadonaga 2003), we found that TATA 8-mer insertion position more than 38 bp upstream of the main TSS increased expression (compared to the background level). However, this increase was higher for upstream insertion positions [-118,-96] (between 88 and 66 bp upstream of the main TSS) and lower for more downstream insertion positions [-95,-69] (between 65 and 39 bp upstream of the main TSS). Interestingly, we also found that shifting the TATA 8-mer had almost no effect on TSS positions, but it did affect TSS usage such that for the more downstream insertion positions [-95,-69] the percent of initiation events at the main TSS decreased and that at alternative downstream TSSs increased (**Fig. 4B**). For the 5 remaining downstream TATA 8-mer insertions at positions -59, -49, -39, -29 and -19, we found that the same TSSs were used and that expression hardly changed. This result demonstrates that TATA elements at the 3' end of promoters are not functional, possibly because the PIC is not recruited to these downstream inserted TATA 8-mers.

To verify that TATA element location affects its function also in native promoters, we performed knockout mutations of 127 TATA elements having the consensus TATAWA (W=A/T) 6-mer within native [-118,-1] promoter regions included in our library (**Fig. 4C**). Compared to the activity of the native core promoter, mutations within the [-118,-99] and [-98,-69] regions greatly reduced expression (median changes of -29.6% and -19.2%, respectively), whereas such mutations had little (if any) effect within the [-68,-1] region (median expression

change of -0.5%). In addition, we performed inversion mutations of 45 TATAAANN (N=A/C/G/T) 8-mers (**Fig. 4D**), and found that, here too, such mutations within the [-118,-99] and [-98,-69] regions greatly reduced expression (median changes of -40.2% and -29.1%, respectively), whereas again, mutations within the [-68,-1] region had little effect on expression (median change of -3.3%). We also found that native TATAWANN 8-mers within the [-118,-69] promoter region are significantly more conserved across yeast species compared to those within the [-68,-1] region ($P<10^{-13}$, **Supplementary Fig. 6A**), and also compared to their own flanking sequences ($P<10^{-6}$, **Supplementary Fig. 6B**), providing further support to the above claims on TATA elements functionality.

Previous studies showed that flanking bases affect in-vivo TF binding specificities, possibly through effects on DNA structure (Gordân et al. 2013; Aow et al. 2013; Rajkumar et al. 2013; Levo et al. 2015). We thus explored the extent to which the immediate context around the TATA element influences its function by additionally randomizing flanking bases around the TATAAAAA and TATATATA 8-mers inserted into position -98. We varied flanking sequences of lengths 2, 5 and 10 bp, and found that such TATA element context variations can induce dramatic effects on the function of the TATA element, even for the short 2 bp flanks (**Fig. 4E**). For example, in the case of TATAAAAA insertion into the *PDC1* background sequence the resulting increase in activity ranged between 24% and 132%, depending on the flanking bases.

Finally, we compared the effects that varying the TATA element sequence had on core promoter activity by inserting different TATA 8-mers into position -98. These included the 8 different 8-mers adhering to the consensus TATAWAWR (W=A/T, R=A/G) (Basehoar et al. 2004), as well as the 144 8-mers that are 1 mismatch away from a consensus 8-mer. We found that with both *PDC1* and *ENO2* backgrounds, consensus TATA 8-mers increased activity significantly more than those with 1 mismatch ($P<10^{-4}$ with both backgrounds, **Fig. 4F**). This result is in agreement with a previous study showing that a consensus TATA element increases expression more than a weaker TATA element (Mogno et al. 2010).

Taken together, our results demonstrate that location, orientation, and flanking bases of TATA elements substantially affect their function, in addition to their sequence specificity.

## Highly expressed core promoters tend to initiate transcription from a narrow region

In a recent computational study we showed an association between focused transcription initiation (transcription initiation within a narrow region around the most frequent TSS) and expression level (Lubliner et al. 2013), relying only on highly expressed constitutive promoters and low resolution TSS distribution data (Miura et al. 2006). Here we sought to test this suggested association, based on our higher resolution TSS distribution and core promoter activity measurements for 252 native [-118,-1] promoter regions. We defined the degree of focused transcription initiation to be the fraction of transcription initiation events that were within 10 bp of the most frequent TSS. Although we found a significant correlation between this measure and core promoter activity ($r$=0.27, $P$<$10^{-4}$, **Fig. 5**), it is clear that the association is not linear, with apparent depletion of core promoters with both high activity and dispersed transcription initiation. We therefore partitioned the native core promoters to those with high activity and to those with lower activity based on the core promoter activity center value (**Fig. 5**, x-axis partitioning), and alternatively partitioned them to those with focused initiation and to those with more dispersed initiation (**Fig. 5**, y-axis partitioning) based on the center value of the degree of focused transcription initiation. Indeed, we found highly active core promoters to be enriched for focused transcription initiation ($P$<0.003, **Fig. 5**), whereas for core promoters with lower activity we found no preference between focused and dispersed transcription initiation (**Fig. 5**).

These results suggest that highly active core promoters are constrained to have transcription initiation focused within a narrow region, while those with lower activity are not and can therefore encode for either focused or dispersed transcription initiation. One explanation for this may be that focused transcription initiation has a small effect on activity, and is therefore optimized only when very high activity is sought. Alternatively, the effect of focused transcription initiation may be irrelevant in lowly active core promoters due to bottlenecks on activity that are further upstream in the core promoter.

## Poly(dA)/poly(dT) orientation at the 3’ end of the promoter affects expression

Poly(dA)/poly(dT) tracts act as nucleosome disfavoring sequences, greatly influencing the nucleosome organization along the DNA (Segal and Widom 2009), thereby affecting the

accessibility of regulatory proteins to the DNA and fine-tuning gene expression regulation (Field et al. 2008; Raveh-Sadka et al. 2012). We performed inversion mutations of 53 poly(dA) and 78 poly(dT) sequences (homopolymeric, at least 6 and 5 bp long respectively) found within native [-118,-1] promoter regions included in our library. Notably, we found that within the [-30,-1] region, inversions of poly(dA) reduced expression (-15.3% median, **Fig. 6A**) and conversely, inversions of poly(dT) increased expression (13.5% median, **Fig. 6B**), demonstrating a substantial role for the orientation of these elements at the 3' end of the core promoter.

## Orthologous RP core promoters of the *Saccharomyces sensu stricto* genus drive similar expression

The RP core promoters of the *Saccharomyces sensu stricto* species tend to be more conserved than upstream promoter regions, yet they too substantially diverged, with some parts showing even less than 70% sequence identity between *S. cerevisiae* and *S. bayanus* (Zeevi et al. 2014). In our library we included the [-118,-1] regions of orthologous RP promoters representing 4 *sensu stricto* species and 67 genes. We found the activity of the *S. cerevisiae* RP core promoters to be highly similar to that of their orthologs from *S. paradoxus* ($r$=0.9, $P<10^{-19}$), *S. mikatae* ($r$=0.84, $P<10^{-15}$) and *S. bayanus* ($r$=0.89, $P<10^{-18}$) (**Fig. 7**), similar to what we recently observed at the level of the entire RP promoter region (Zeevi et al. 2014). These results suggest that in order for orthologous *sensu stricto* RP promoters to maintain their promoter activity during evolution they had to maintain the activity of their core promoter, and the alternative option whereby their core promoter activity might have diverged and be compensated for by other mechanisms is ruled out. This is also consistent with the high correlation that we observed between their core promoter activity and that of their entire promoter (**Fig. 2A**).

## Discussion

In summary, here we presented the first comprehensive study of yeast core promoter sequences, providing massively parallel measurements of both core promoter activity and TSS distributions for thousands of native and synthetic core promoters. We found core promoter activity to be highly correlated to the activity of the entire promoter (**Fig. 2**), demonstrating that core promoter sequence is a major determinant of promoter activity.

Our results are in line with the well established role played by TF binding sites and chromatin in determining promoter activity (for a recent review see (Levo and Segal 2014)). The core promoter sequence is another regulatory layer that affects transcription, and our results demonstrate that the contribution of this layer to determining promoter output is as important as that of the other, better studied, regulatory layers. Thus, optimizations of TF binding, the chromatin landscape (and its effect on TF binding and PIC recruitment) and the core promoter are all necessary for high promoter activity, and consequently it may suffice to de-optimize just one of them to introduce a bottleneck into transcription initiation and reduce promoter activity.

Recent studies showed that expression can be fine-tuned through the effects of flanking bases on TF binding site affinities or through manipulations of nucleosome disfavoring sequences (e.g. poly(dA:dT) tracts) (Raveh-Sadka et al. 2012; Rajkumar et al. 2013). Here, we found that sequence variation within different core promoter regions resulted in a wide and continuous spectrum of core promoter activities (**Fig. 3)**. We also found that modifying the sequence, location and flanking bases of TATA elements greatly affects core promoter activity (**Fig. 4**). Hence, fine-tuning expression can also be achieved through variation of the core promoter sequence.

We also showed that highly active core promoters tend to have transcription initiation focused within a narrow region around the main TSS, while lower activity core promoters may have both focused or dispersed transcription initiation (**Fig. 5**), suggesting that TSS distribution has an effect on core promoter activity.

For poly(dA)/poly(dT), we demonstrated that their orientation affects expression at the 3' end of promoters (**Fig. 6**). Our results concur with past computational studies suggesting that higher A-content at and downstream of the main TSS contributes to higher expression levels (Maicas

and Friesen 1990; Lubliner et al. 2013), and are a demonstration of how one biological sequence signal can introduce biases into another that are not necessarily relevant to the biological function of the latter. To encode for nucleosome depletion at its 3' end the promoter can include a poly(dA) or poly(dT) sequence there, but the selection between the two alternatives is biased by the role played by A-content in affecting expression, probably through transcription initiation (Zhang and Dietrich 2005; Lubliner et al. 2013).

Along these lines, we also suggest an alternative explanation to observations made in a recent study of yeast promoters, reporting a bias of poly(dT) upstream vs. poly(dA) downstream of position -75 relative to the (main) TSS (Wu and Li 2010). This bias was suggested to underlie the determination of the nucleosome free region (NFR) center. However, the fact that most promoters do not actually contain both poly(dT) and poly(dA) (Wu and Li 2010) does not support this. Rather, we suggest that in promoters lacking a consensus TATA element there is a bias in favor of higher A-content at short regions acting as clusters of weak TATA elements (Sugihara et al. 2011; Rhee and Pugh 2012), which adds to the constraint to encode for nucleosome depletion through poly(dT) or poly(dA) tracts.

We also examined the evolution of orthologous RP core promoters in the *Saccharomyces sensu stricto* genus. In a previous study we showed that *sensu stricto* orthologous RP promoters are highly diverged in sequence yet their promoter activities were conserved (Zeevi et al. 2014). Consistently, here we found that the activities of orthologous *sensu stricto* RP core promoters were conserved (**Fig. 7**). Since these orthologous core promoters substantially diverged in sequence (Zeevi et al. 2014) this could be achieved either through robustness to sequence variation or, as we previously demonstrated for the *RPL4A* and the *RPL5* genes (Zeevi et al. 2014), through compensatory sequence variation within the core promoter.

A recent study suggested that the GAAAA 5-mer is a conserved yeast promoter element, functioning as a TBP binding site in promoters lacking a consensus TATA element (Seizl et al. 2011). We performed knockout mutations of 122 GAAAA 5-mers within native [-118,-1] promoter regions included in our library, and found these mutations to have little effect on expression in both the [-118,-99] and the [-98,-59] regions (**Supplementary Fig. 7**), providing evidence against the GAAAA 5-mer being a distinct element with substantial functional importance. More likely, this 5-mer may sometimes be part of several redundant weak TATA

elements that have a few mismatches to the TATA element consensus (Sugihara et al. 2011; Rhee and Pugh 2012).

Although we demonstrated the crucial role of the core promoter in determining transcription levels in yeast, we expect our findings to apply beyond yeast. Recent studies of metazoan core promoters support this assertion (Lubliner et al. 2013; Zehavi et al. 2014). We thus expect our study to encourage similar studies of human and other metazoan core promoters, seeking to better understand various aspects such as the extent to which core promoters determine and tune transcription levels, and the rules and constraints on their evolution.

# Methods

## Library construction and expression measurements

Library construction and expression measurements were conducted as previously described (Sharon et al. 2012), except for the following changes.

Sequences were designed *in silico* and synthesized as 200 bases long ssDNA oligos by Agilent Technologies (LeProust et al. 2010). Most oligo design details appear above in the Results section. We also note, as shown in **Figure 1**, that the 54 bp long 3' end of each oligo included the 54 bp long prefix of the *YFP* reporter.

Library sequences were amplified using two sequential PCR reactions. First reaction: 95ºC for 3 min and then 7 cycles of 95ºC for 30 sec and 68ºC for 1 min, each, and finally one cycle at 68ºC for 5 min. Second reaction: 95ºC for 3 min, 9 cycles of 95ºC for 30 sec, 50ºC for 30 sec and 72ºC for 30 sec, each, and finally one cycle of 68ºC for 5 min. The 18 bp universal primers sequences used to amplify the library were: forward primer- 5'-TTGTACCTGGTCTCTGCG-3' and reverse primer- 5'-TAATTCCACCAAAATGGG-3'. The amplified library was cut with SexAI and BstXI restriction enzymes (Fermentas) and then ligated into our pCore plasmid (see **Supplementary Fig. 8**). The ligation products were transformed into E. cloni electrocompetent cells (Lucigen), which were then plated on LB/ampicillin plates. The transformation yielded 1.6 million colonies that were scraped form the plates and the plasmids were purified using a plasmid maxi kit (Qiagen).

Next, the plasmid library was transformed into yeast cells (strain Y8205) by electrotporation and the transformants were grown on liquid SCD-URA selective medium. The culture was grown to saturation and then regrown for sorting by diluting 1:500 in the same medium and growing to mid-exponential phase. Cells were sorted (BD FACSAria III) into 16 bins according to their YFP/mCherry values. We sorted only cells that were filtered to have relatively homogeneous size and mCherry fluorescence (corresponding to ~1-2 plasmid copies).

Following sorting, cells were grown in 6ml SCD-URA medium to stationary phase. One million cells from each bin were taken for colony PCR, using the following primers. The 3' primer was common to all bins: 5'-NNNNNGAATAATTCTTCACCTTTAG-3' (N = random

nucleotide). The 5' primer had a common sequence along with a unique 5 bp long upstream barcode sequence specific to each bin (represented by XXXXX): 5'-XXXXXTTGTACCTGGTCTCTGCG-3'. The PCR output was then taken for parallel sequencing (Illumina HiSeq 2000).

## 5' rapid amplification of cDNA ends (5'RACE)

Yeast cells were grown to stationary phase and regrown to mid-exponential phase in SCD-URA medium (as was done prior to expression measurements by FACS, see above). Cells were aliquoted to 25ml and centrifuged to pellet cells at 3000g for 8min. The growth medium was removed and total RNA was extracted using lyticase digestion followed by TriReagent (MRC) RNA extraction protocol.

RNA libraries for transcription start site mapping (5'-end RNA-seq) were prepared as in (Wurtzel et al. 2010). In brief, RNA was incubated with Tabacco Acid Pyrophosphatase (TAP, Epicentre) to treat 5' ends, and 3' ends were blocked using NaOI4. Illumina's 5' adapter was ligated to the RNA with T4 RNA ligase (NEB). cDNA priming was done using a *YFP* gene specific primer (GSP). Following cDNA synthesis, *YFP* amplicons were amplified 18 cycles using a nested *YFP* GSP attached to an Illumina 3' adapter and a 5' Illumina adapter as forward primer.

## Learning linear models

For each of the 4 mutation subsets (see above) we computed base content and *k*-mer count features for both native and mutated sequences. In the case of the mutated functional regions (PIC, Scanning, Initiation) the features were computed over the entire region. In the case of the sliding window mutations, features were computed over various windows along the core promoter region. Then, from each feature we computed a mutational sequence difference feature by subtracting from each value of a mutated sequence the value of the respective native sequence. In addition, for each mutated sequence its effect on expression was computed to be the percent change in core promoter activity from that of the respective native sequence.

We learned linear models that predict mutation effects on expression from mutational sequence difference features. Model learning was performed using a *K*-fold cross validation scheme (*K*=5 for the functional region mutations, *K*=10 for the sliding window mutations), with models learned on training data and their performance assessed on held out test data. A complete description of the cross validated linear model learning scheme appears in the **Supplementary Note**. We repeated the learning several times, starting with an initial set containing only base content features, and sequentially adding higher order *k*-mers (for *k*=2,…,5) to the initial set. We reported the results attained when the mean test $R^2$ was highest. For the PIC region this was with *k*=4, for the Scanning and Initiation regions this was with *k*=2 and for the sliding window mutations this was with *k*=5 (**Supplementary Figs. 1-4**). The *K*-fold cross validation scheme produced *K* models, and we reported the features that were included in at least 60% of them (**Supplementary Figs. 1-4**).

## Data Access

Raw and processed expression data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE68168.

## Acknowledgements

We thank Zohar Yakhini and Paige Anderson from Agilent Technologies for their role and assistance in producing the DNA oligos, Eilon Sharon and Leeat Keren for their help with performing FACS measurements, Tali Avnit-Sagi for her help with running the MiSeq sequencing of the 5'RACE products and also Shira Weingarten-Gabbay and Michal Levo for valuable discussions. This work was supported by grants from the European Research Council (ERC), the US National Institutes of Health (NIH), and the Israel Science Foundation (ISF) to E. Segal.

## References

Aow JSZ, Xue X, Run J-Q, Lim GFS, Goh WS, Clarke ND. 2013. Differential binding of the related transcription factors Pho4 and Cbf1 can tune the sensitivity of promoters to different levels of an induction signal. *Nucleic Acids Res* **41**: 4877–87.

Basehoar AD, Zanton SJ, Pugh BF. 2004. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116**: 699–709.

Bushnell DA, Westover KD, Davis RE, Kornberg RD. 2004. Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms. *Science* **303**: 983–8.

Chen W, Struhl K. 1988. Saturation mutagenesis of a yeast his3 "TATA element": genetic evidence for a specific TATA-binding protein. *Proc Natl Acad Sci U S A* **85**: 2691–5.

Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, Segal E. 2013. Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proc Natl Acad Sci U S A* **110**: E2792–801.

Faitar SL, Brodie SA, Ponticelli AS. 2001. Promoter-specific shifts in transcription initiation conferred by yeast TFIIB mutations are determined by the sequence in the immediate vicinity of the start sites. *Mol Cell Biol* **21**: 4427–40.

Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, Lubling Y, Widom J, Segal E. 2008. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput Biol* **4**: e1000216.

Fishburn J, Hahn S. 2012. Architecture of the yeast RNA polymerase II open complex and regulation of activity by TFIIF. *Mol Cell Biol* **32**: 12–25.

Furter-Graves EM, Hall BD. 1990. DNA sequence elements required for transcription initiation of the Schizosaccharomyces pombe ADH gene in Saccharomyces cerevisiae. *Mol Gen Genet* **223**: 407–16.

Giardina C, Lis JT. 1993. DNA melting on yeast RNA polymerase II promoters. *Science* **261**: 759–62.

Goel S, Krishnamurthy S, Hampsey M. 2012. Mechanism of start site selection by RNA polymerase II: interplay between TFIIB and Ssl2/XPB helicase subunit of TFIIH. *J Biol Chem* **287**: 557–67.

Gordân R, Shen N, Dror I, Zhou T, Horton J, Rohs R, Bulyk ML. 2013. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep* **3**: 1093–104.

Hahn S, Hoar ET, Guarente L. 1985. Each of three "TATA elements" specifies a subset of the transcription initiation sites at the CYC-1 promoter of Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **82**: 8562–6.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**: 362–366.

Keren L, Zackay O, Lotan-Pompan M, Barenholz U, Dekel E, Sasson V, Aidelberg G, Bren A, Zeevi D, Weinberger A, et al. 2013. Promoters maintain their relative activity levels under different growth conditions. *Mol Syst Biol* **9**: 701.

Kuehner JN, Brow DA. 2006. Quantitative analysis of in vivo initiator selection by yeast RNA polymerase II supports a scanning model. *J Biol Chem* **281**: 14119–28.

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 Res* **38**: 2522–40.

Levo M, Segal E. 2014. In pursuit of design principles of regulatory sequences. *Nat Rev Genet* **15**: 453–68.

Levo M, Zalckvar E, Sharon E, Dantas Machado AC, Kalma Y, Lotam-Pompan M, Weinberger A, Yakhini Z, Rohs R, Segal E. 2015. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res* gr.185033.114.

Lieb JD, Liu X, Botstein D, Brown PO. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* **28**: 327–34.

Lubliner S, Keren L, Segal E. 2013. Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Res* **41**: 5569–81.

Mahadevan S, Struhl K. 1990. Tc, an unusual promoter element required for constitutive transcription of the yeast HIS3 gene. *Mol Cell Biol* **10**: 4447–55.

Maicas E, Friesen JD. 1990. A sequence pattern that occurs at the transcription initiation region of yeast RNA polymerase II promoters. *Nucleic Acids Res* **18**: 3387–93.

Miller G, Hahn S. 2006. A DNA-tethered cleavage probe reveals the path for promoter DNA in the yeast preinitiation complex. *Nat Struct Mol Biol* **13**: 603–10.

Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, Ito T. 2006. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci U S A* **103**: 17846–51.

Mogno I, Vallania F, Mitra RD, Cohen BA. 2010. TATA is a modular component of synthetic promoters. *Genome Res* **20**: 1391–7.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344–1349.

Rajkumar AS, Dénervaud N, Maerkl SJ. 2013. Mapping the fine structure of a eukaryotic promoter input-output function. *Nat Genet* **45**: 1207–15.

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. *Nat Genet* **44**: 743–50.

Rhee HS, Pugh BF. 2012. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**: 295–301.

Segal E, Widom J. 2009. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struct Biol* **19**: 65–71.

Seizl M, Hartmann H, Hoeg F, Kurth F, Martin DE, Söding J, Cramer P. 2011. A conserved GA element in TATA-less RNA polymerase II promoters. *PLoS One* **6**: e27595.

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. *Nat Biotechnol* **30**: 521–30.

Singer VL, Wobbe CR, Struhl K. 1990. A wide variety of DNA sequences can functionally replace a yeast TATA element for transcriptional activation. *Genes Dev* **4**: 636–45.

Smale ST, Kadonaga JT. 2003. The RNA polymerase II core promoter. *Annu Rev Biochem* **72**: 449–79.

Stewart JJ, Stargell LA. 2001. The stability of the TFIIA-TBP-DNA complex is dependent on the sequence of the TATAAA element. *J Biol Chem* **276**: 30078–84.

Sugihara F, Kasahara K, Kokubo T. 2011. Highly redundant function of multiple AT-rich sequences as core promoter elements in the TATA-less RPS5 promoter of Saccharomyces cerevisiae. *Nucleic Acids Res* **39**: 59–75.

Tanay A, Regev A, Shamir R. 2005. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci U S A* **102**: 7203–8.

Wu R, Li H. 2010. Positioned and G/C-capped poly(dA:dT) tracts associate with the centers of nucleosome-free regions in yeast promoters. *Genome Res* **20**: 473–84.

Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R. 2010. A single-base resolution map of an archaeal transcriptome. *Genome Res* **20**: 133–41.

Zeevi D, Lubliner S, Lotan-Pompan M, Hodis E, Vesterman R, Weinberger A, Segal E. 2014. Molecular dissection of the genetic mechanisms that underlie expression conservation in orthologous yeast ribosomal promoters. *Genome Res* **24**: 1991–1999.

Zehavi Y, Kuznetsov O, Ovadia-Shochat A, Juven-Gershon T. 2014. Core promoter functions in the regulation of gene expression of Drosophila dorsal target genes. *J Biol Chem* **289**: 11993–2004.

Zhang Z, Dietrich FS. 2005. Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE. *Nucleic Acids Res* **33**: 2838–51.

## Figure Legends

### Figure 1

**Illustration of our experimental system.** Oligonucleotides from a library comprising 13,000 designed synthetic sequences (Agilent Technologies, Santa Clara, CA) containing a 118 bp long variable core promoter sequence were ligated into a low copy plasmid (top). Designed sequences included 7,536 unique core promoter sequences, and for 5,464 of them we designed a second instance that was additionally barcoded by synonymous mutations within the first 36 bp of the *YFP*. The barcoded sequences also differed from the non-barcoded ones by 4 mismatches within the 10 bp upstream of the YFP. The plasmid pool was transformed into yeast to create a heterogeneous pool of yeast cells, each cell expressing YFP at a different level (middle). To measure expression, cells were sorted using fluorescence activated sorting (FACS) into 16 expression bins by their YFP/mCherry ratio, and the core promoter sequences were amplified using bin-specific barcoded primers and sent to parallel sequencing (left pipeline). Sequencing reads coming from YFP-barcoded instances were removed. Each read was then mapped to a YFP/mCherry bin and a core promoter sequence. This gave for each core promoter sequence the binned distribution of YFP/mCherry levels over the cells that had that sequence (bottom left), from which we computed the mean YFP/mCherry (see Supplementary Note). To map TSSs, we extracted total RNA from the pool of yeast cells, performed 5'RACE using primers specific to the *YFP* sequence, and sequenced the products (right pipeline). Sequencing reads not mapping to YFP-barcoded instances were removed. Each read was then mapped to a core promoter sequence by its YFP barcode, enabling us to compute the transcription initiation landscape of YFP-barcoded core promoter sequences (bottom right, see Supplementary Note).

### Figure 2

**Core promoter activity is highly correlated to the activity of the entire promoter.** (A) A comparison of our core promoter activity measurements (x-axis) to previously measured promoter activities (y-axis) for 238 *Saccharomyces sensu stricto* RP promoters (Zeevi et al. 2014) reveals a high correlation between the two measures. Dark red dots mark promoters that

are not regulated by Rap1. (B) Similar to (A) for 133 constitutively expressed *S. cerevisiae* genes (Keren et al. 2013).

**Figure 3**

**Sequence variation in different core promoter regions substantially affects activity.** Results of learning linear models that predict the effects of mutating various native core promoter regions (see main text) on core promoter activity, based on sequence features that measure differences between mutant and native core promoters. We used a *K*-fold cross validation scheme, such that each mutant appeared once in a held-out test set, and *K*-1 times in a training set. (A) Results for PIC region mutations. For each mutated core promoter we plotted its measured percent change (compared to the native core promoter) in core promoter activity (x-axis) against its predicted one (y-axis, predicted by the linear model learned when that mutant was part of the held-out test set). Dotted grey lines mark the axes zero values. We also report mean performance measures (*r* and $R^2$ statistics) of the models over the test sets. (B) Same as (A) for Scanning region mutations. (C) Same as (A) for Initiation region mutations. (D) Same as (A) for sliding window mutations. (E) An illustration summarizing classes of sequence features included into our linear models and their predicted effect on core promoter activity. All learned features are specified in Supplementary Figures 1-4. The golden right arrow marks the translation start site.

**Figure 4**

**TATA element location, orientation, sequence and flanking bases affect its functionality.** (A) We inserted the TATA consensus 8-mer TATAAAAA into different positions along the *PDC1* background sequence (see main text). Each row in the left panel heatmap corresponds to one insertion case, with TATA start position marked in dark blue (in a few cases the insertion actually resulted in two overlapping TATA 8-mers, and then both start positions are marked), and the measured TSS distribution appears in red and yellow colors (see color bar on the left). The effect of every insertion on core promoter activity is shown by the corresponding bar within

the right panel. The light blue dashed lines separate the results of three regions: [-118,-96], [-95,-69] and the 5 insertion positions further downstream. Note that there are a few instances with missing TSS or expression data. (B) For the same data shown in (A), an illustration of the percent initiation events (y-axis) at position -30 (red dots) vs. positions [-29,-1] (blue dots), as a function of the TATAAAAA 8-mer insertion start position (x-axis). The trend lines of corresponding colors show the moving average using a sliding window of length 10. (C) Box plots of the percent changes to core promoter activity caused by knockout mutations of native TATA elements having the consensus 6-mer TATAWA (W=A/T), in three core promoter windows: [-118,-99], [-98,-69] and [-68,-1]. Assignment to windows was based on the TATAWA start position. (D) Same as in (C) for inversions of native TATAAANN (N=A/C/G/T) 8-mers. (E) For insertions of the TATAAAAA and TATATATA consensus TATA 8-mers into position -98 of the *ENO2* and *PDC1* backgrounds we also generated instances in which we additionally randomized their flanking sequences of lengths 2, 5 or 10 bp. For each such instance we plotted the percent change to core promoter activity (y-axis). Pink dots mark cases with random flanks of 2 bp, red dots – 5 bp, and dark red dots – 10 bp. Dashed light blue lines mark the value measured for insertions without flanking sequence randomization. (F) Box plots of the percent changes to core promoter induced expression caused by insertion (into position -98) of either consensus TATAWAWR (W=A/T, R=A/G) 8-mers or TATA 8-mers that are 1 mismatch away from a consensus 8-mer. Top (bottom) panel is for insertions into the *PDC1* (*ENO2*) background.

**Figure 5**

**Highly expressed core promoters tend to have focused transcription initiation.** A comparison of our core promoter activity measurements (x-axis) to our measure of focused transcription initiation (y-axis) for 252 native core promoters (taken from the union of the two sets in Fig. 2) reveals a significant correlation between the two measures. Based on the center values ((min.+max.)/2) of the two measures, native core promoters were classified as having either high or low activity (x-axis classification) and as having either focused or dispersed transcription initiation (y-axis classification). Core promoters with high activity were found to be enriched for focused transcription initiation.

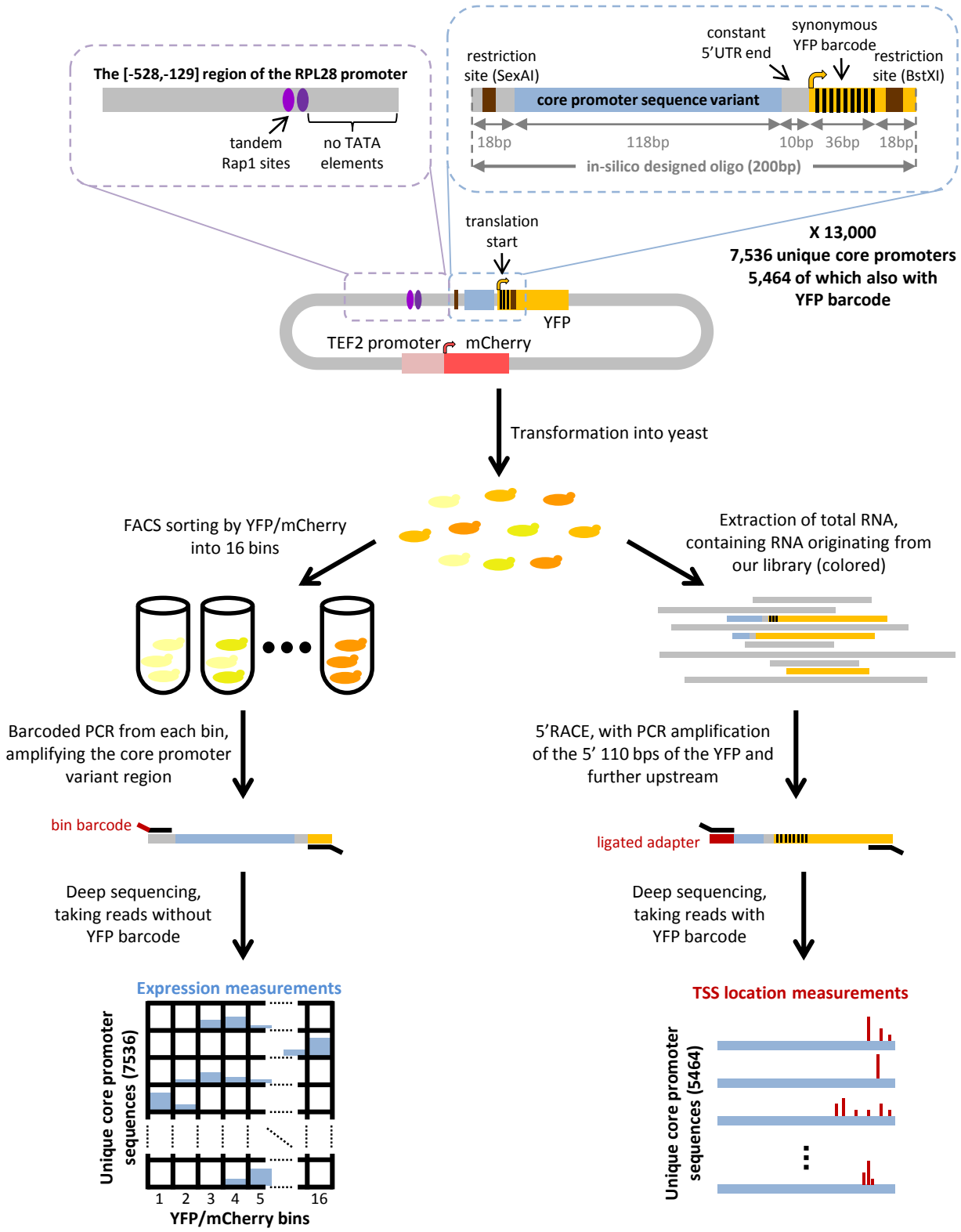
**Figure 6**

**Effect of poly(dA)/poly(dT) inversions in different parts of the core promoter.** (A) Box plots of the percent changes to core promoter activity caused by inversions of native poly(dA) homopolymers in four core promoter windows: [-118,-91], [-90,-61], [-60,-31] and [-30,-1]. Assignment to windows was based on the poly(dA) start position. (B) Same as in (A) for inversions of native poly(dT) homopolymers.

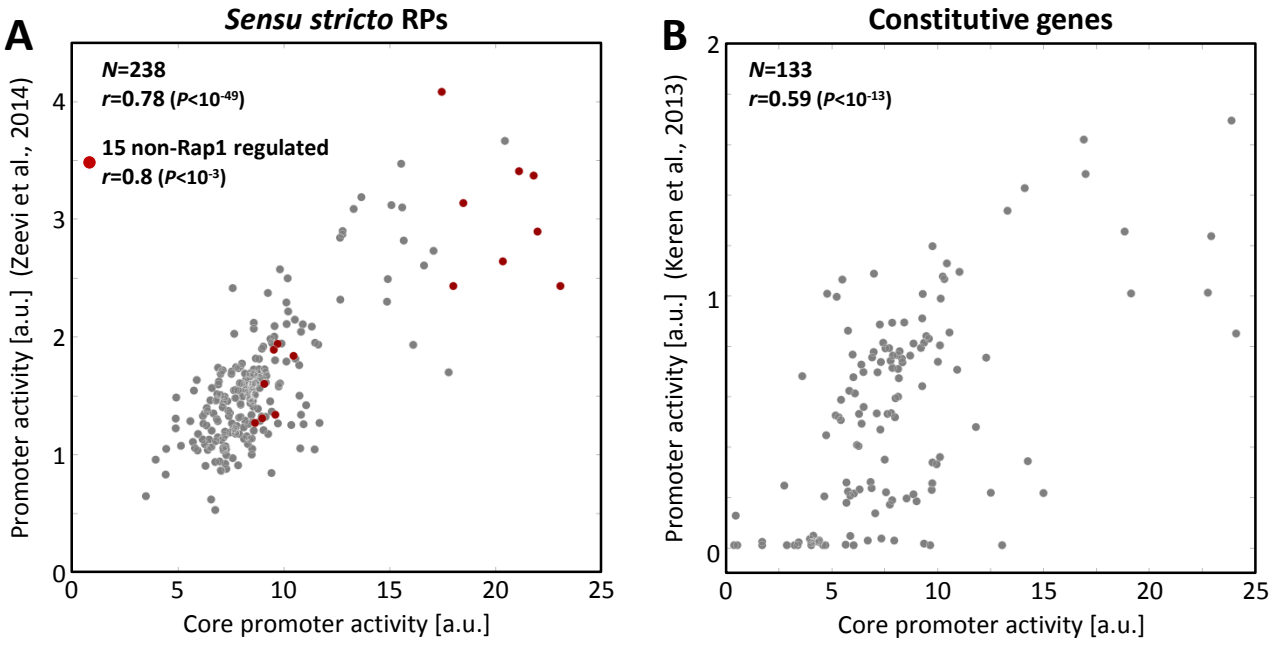
**Figure 7**

**Core promoter activity is conserved between orthologous *Saccharomyces sensu stricto* RPs.** A comparison of the core promoter activity of native RP core promoters from *S. cerevisiae* and their orthologous counterparts from each of *S. paradoxus*, *S. mikatae* and *S. bayanus* reveals very high correlations.

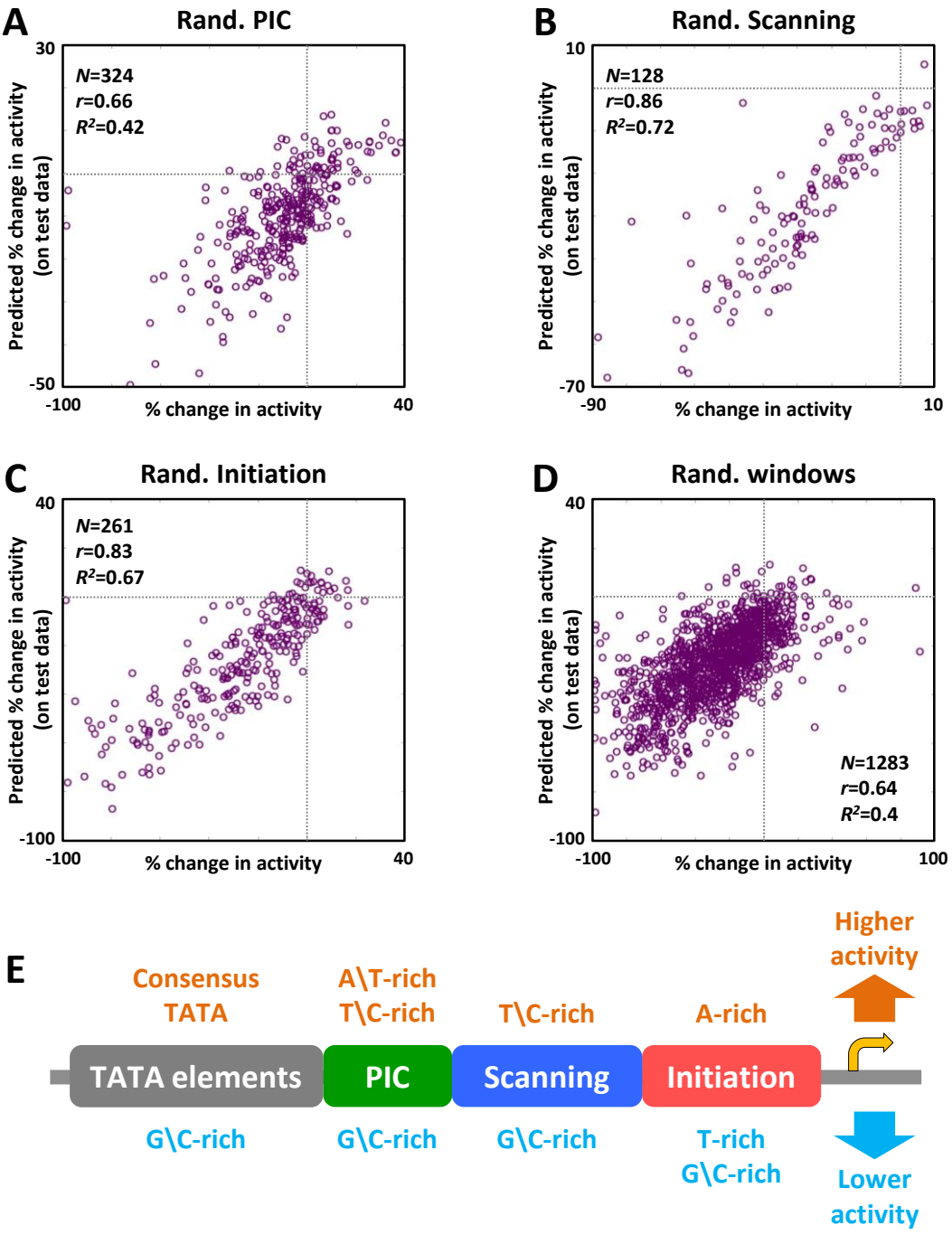
# Figure 1


The [-528,-129] region of the RPL28 promoter
tandem Rap1 sites
no TATA elements
restriction site (SexAI)
constant 5'UTR end
synonymous YFP barcode
restriction site (BstXI)
core promoter sequence variant
18bp
118bp
10bp
36bp
18bp
in-silico designed oligo (200bp)
translation start
X 13,000
7,536 unique core promoters
5,464 of which also with YFP barcode
YFP
TEF2 promoter
mCherry
Transformation into yeast
FACS sorting by YFP/mCherry into 16 bins
Extraction of total RNA, containing RNA originating from our library (colored)
Barcoded PCR from each bin, amplifying the core promoter variant region
5'RACE, with PCR amplification of the 5' 110 bps of the YFP and further upstream
bin barcode
ligated adapter
Deep sequencing, taking reads without YFP barcode
Deep sequencing, taking reads with YFP barcode
Expression measurements
Unique core promoter sequences (7536)
1 2 3 4 5 16
YFP/mCherry bins
TSS location measurements
Unique core promoter sequences (5464)

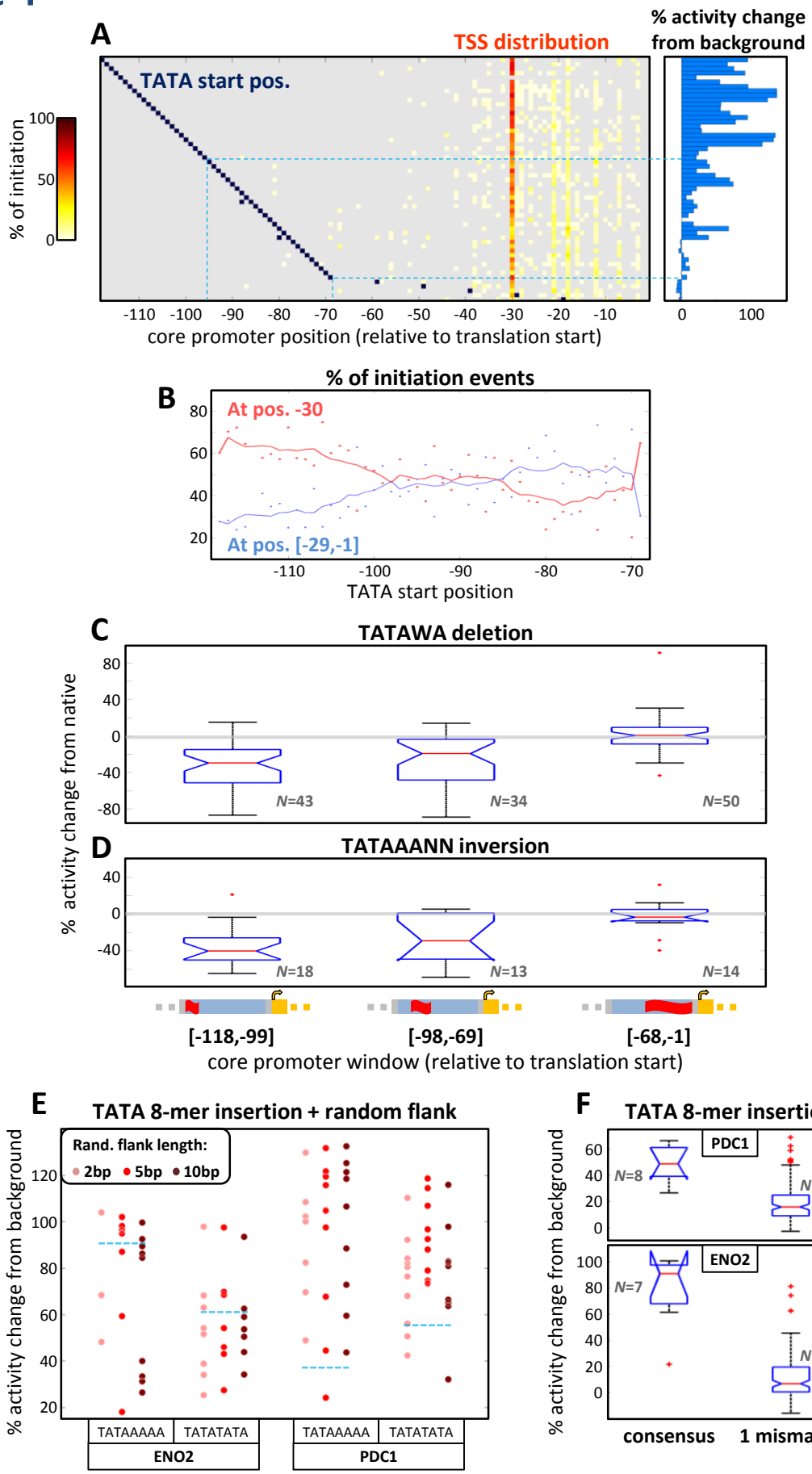
# Figure 2


A
Sensu stricto RPs
N=238
r=0.78 (P<10-49)
15 non-Rap1 regulated
r=0.8 (P<10-3)
Promoter activity [a.u.] (Zeevi et al., 2014)
Core promoter activity [a.u.]
0
1
2
3
4
5
10
15
20
25
B
Constitutive genes
N=133
r=0.59 (P<10-13)
Promoter activity [a.u.] (Keren et al., 2013)
Core promoter activity [a.u.]
0
1
2
5
10
15
20
25

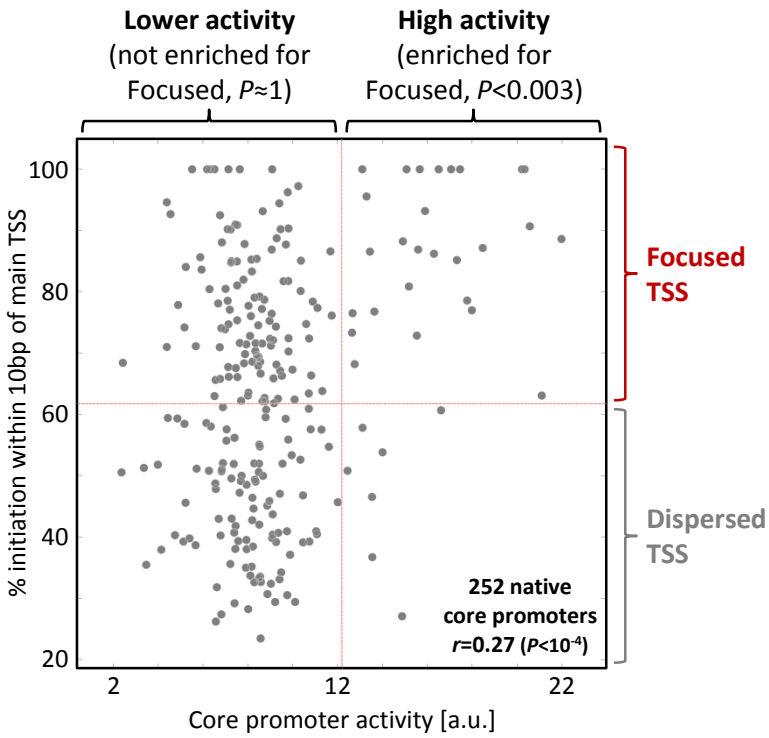
# Figure 3


A
Rand. PIC
N=324
r=0.66
R²=0.42
Predicted % change in activity (on test data)
% change in activity
30
-50
-100
40
B
Rand. Scanning
N=128
r=0.86
R²=0.72
10
-70
-90
C
Rand. Initiation
N=261
r=0.83
R²=0.67
D
Rand. windows
N=1283
r=0.64
R²=0.4
100
E
Consensus TATA
A\T-rich T\C-rich
T\C-rich
A-rich
Higher activity
TATA elements
PIC
Scanning
Initiation
G\C-rich
G\C-rich
G\C-rich
T-rich G\C-rich
Lower activity

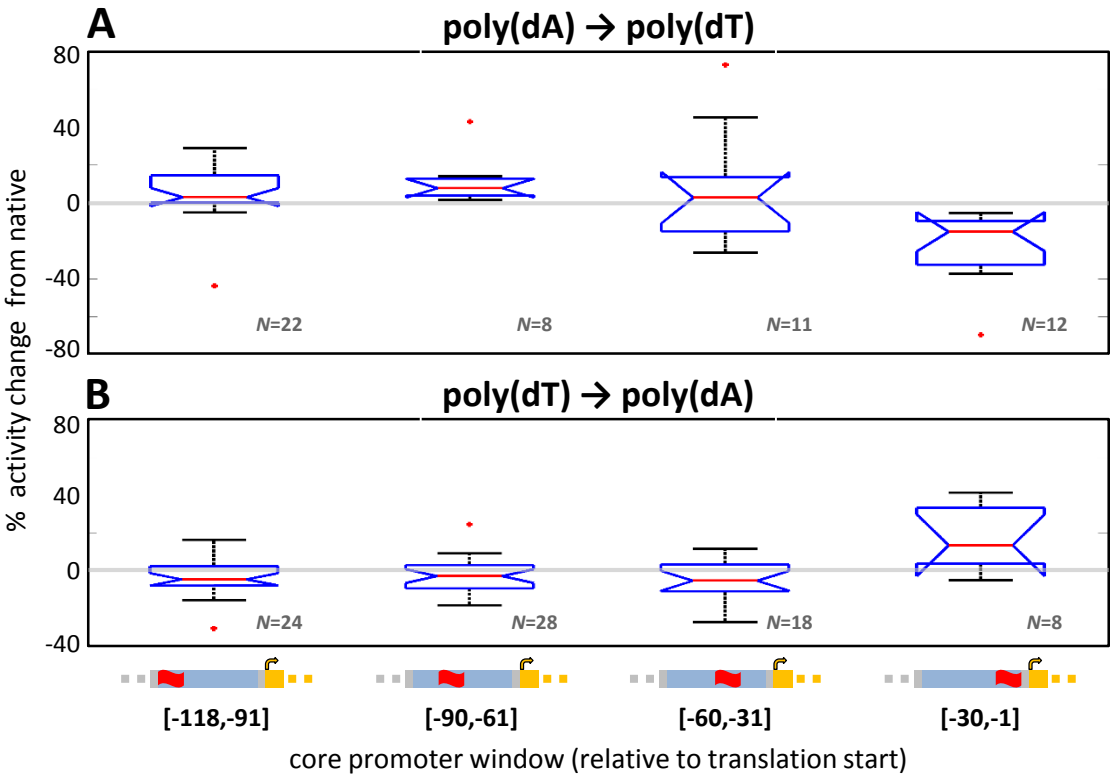
# Figure 4


A
TSS distribution
% activity change from background
TATA start pos.
% of initiation
100
50
0
-110 -100 -90 -80 -70 -60 -50 -40 -30 -20 -10
0 100
core promoter position (relative to translation start)
B
% of initiation events
At pos. -30
At pos. [-29,-1]
80
60
40
20
-110 -100 -90 -80 -70
TATA start position
C
TATAWA deletion
% activity change from native
80
40
0
-40
-80
N=43
N=34
N=50
D
TATAAANN inversion
40
0
-40
N=18
N=13
N=14
[-118,-99]
[-98,-69]
[-68,-1]
core promoter window (relative to translation start)
E
TATA 8-mer insertion + random flank
% activity change from background
Rand. flank length:
2bp 5bp 10bp
120
100
80
60
40
20
TATAAAAA TATATATA TATAAAAA TATATATA
ENO2
PDC1
F
TATA 8-mer insertion
% activity change from background
PDC1
N=8
N=141
60
40
20
0
ENO2
N=7
N=134
100
80
60
40
20
0
consensus
1 mismatch

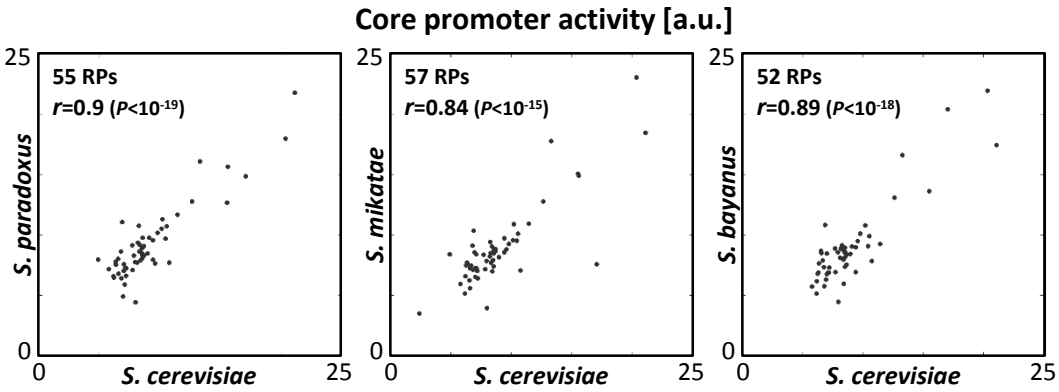
# Figure 5


Lower activity
(not enriched for
Focused, P≈1)
High activity
(enriched for
Focused, P<0.003)
Focused
TSS
Dispersed
TSS
% initiation within 10bp of main TSS
100
80
60
40
20
2
12
22
Core promoter activity [a.u.]
252 native
core promoters
r=0.27 (P<10-4)

# Figure 6


A
poly(dA) → poly(dT)
80
40
0
-40
-80
% activity change from native
N=22
N=8
N=11
N=12
B
poly(dT) → poly(dA)
80
40
0
-40
N=24
N=28
N=18
N=8
[-118,-91]
[-90,-61]
[-60,-31]
[-30,-1]
core promoter window (relative to translation start)

# Figure 7


Core promoter activity [a.u.]
55 RPs
r=0.9 (P<10-19)
S. paradoxus
S. cerevisiae
57 RPs
r=0.84 (P<10-15)
S. mikatae
S. cerevisiae
52 RPs
r=0.89 (P<10-18)
S. bayanus
S. cerevisiae
0
25

# Core promoter sequence in yeast is a major determinant of expression level

Shai Lubliner, Ifat Regev, Maya Lotan-Pompan, et al.

*Genome Res.* published online May 12, 2015
Access the most recent version at doi:10.1101/gr.188193.114