Quantitative analysis of transcription start site selection in *Saccharomyces cerevisiae* determines contributions of DNA sequence and RNA Polymerase II activity

Yunye Zhu[1], Irina O. Vvedenskaya[2], Bryce E. Nickels[2] and Craig D. Kaplan[1]

[1] Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA
[2] Department of Genetics and Waksman Institute, Rutgers University, Piscataway, NJ 08854, USA

Abstract

DNA sequence at Transcription Start Sites (TSSs) is a key determinant of initiation by RNA Polymerase II (Pol II). To function as a TSS, an initiation compatible sequence must be specified by a promoter in an appropriate chromatin context. We report the development of a method for quantitative analysis of transcription initiation by Pol II that involves construction of DNA libraries of barcoded promoter variants, production of RNA transcripts, and analysis of transcript 5' ends and transcript yields (Pol II MAssively Systematic Transcript End Readout, "Pol II MASTER"). Using Pol II MASTER, we measure the efficiency of transcription initiation during "promoter scanning" by *Saccharomyces cerevisiae* Pol II for ~1 million unique TSS sequences. Furthermore, we employ Pol II MASTER to determine how Pol II activity, known to widely alter TSS selection *in vivo*, alters TSS efficiencies across our promoter variants. Pol II MASTER recapitulates known critical qualities of *Saccharomyces cerevisiae* TSS -8, -1, and +1 positions while demonstrating that surrounding sequences modulate initiation efficiency over a wide range. We discover functional interactions between neighboring sequence positions, indicating that adjacent positions likely function together. We demonstrate that initiation efficiencies are altered for +1 A TSSs relative to +1 G TSSs when Pol II activity is perturbed through genetic means. Pol II MASTER provides data for predictive models of TSS initiation efficiency at genomic promoters.

Introduction

In transcription initiation, RNA polymerase II (Pol II) binds promoter DNA through interactions with core promoter elements, unwinds a turn of promoter DNA forming a Pol II-promoter open complex containing a single-stranded "transcription bubble", and selects a promoter position within a region competent for initiation to serve as the transcription start site (TSS). At the majority of Pol II promoters in eukaryotes, TSS selection occurs at multiple positions (Carninci et al., 2005; Chen et al., 2013; Chia et al., 2021; Consortium et al., 2014; Hoskins et al., 2011; Nepal et al., 2013; Park, Morris, Battenhouse, & Iyer, 2014; Pelechano, Wei, & Steinmetz, 2013; Yamashita et al., 2011; Z. Zhang & Dietrich, 2005; Zheng et al., 2011). Thus, the overall rate of gene expression at the majority of Pol II promoters is determined by the efficiency of initiation from several distinct TSS positions. In addition, studies have suggested that alternative TSS selection can lead to differences in mRNA features, translation activity and subsequential protein levels and functions (Cheng et al., 2018; Pelechano et al., 2013; Rojas-Duran & Gilbert, 2012), and therefore is widespread in different cell types (Consortium et al., 2014), developmental processes (Batut, Dobin, Plessy, Carninci, & Gingeras, 2013; Cheng et al., 2018; Chia et al., 2021; P. Zhang et al., 2017), growth conditions (Lu & Lin, 2019), responses to environmental changes, and cancers (Boyd et al., 2018; Demircioglu et al., 2019; Thorsen et al., 2011).

Pol II initiation in yeast proceeds by a promoter scanning mechanism, where the Pol II Pre-initiation Complex (PIC), comprising Pol II and initiation factors, assembles upstream of the initiation region and then scans downstream to select a TSS position (Fazal, Meng, Murakami, Kornberg, & Block, 2015; Hampsey, 1998; Kaplan, Jin, Zhang, & Belyanin, 2012; Kuehner & Brow, 2006; Lis, 1993; Miller & Hahn, 2006; Chenxi Qiu et al., 2020; Zhao et al., 2021). The efficiency of initiation at a given position within a region allowing initiation depends on its location relative to the core promoter region, with bases scanned from upstream to downstream from the core promoter, and on DNA sequence, with the template base specifying the TSS position and position immediately upstream of the TSS (positions +1 and -1, respectively) making the largest contributions. In particular, there is a strong preference for an R:Y base pair at position +1 and Y:R base pair at position -1 (reflected as a $Y_{-1}R_{+1}$ “initiator” sequence on the coding strand; Y=pYrimidine; R=puRine). Furthermore, there is a near universal preference for transcription initiation by multi-subunit and single-subunit RNAPs in all domains of life (Bucher, 1990; Carninci et al., 2006; Chen et al., 2013; Corden et al., 1980; Cortes et al., 2013; Furter-Graves & Hall, 1990; Hahn S, 1985; Hashimoto et al., 2004; Healy, Helser, & Zitomer, 1987; Kim et al., 2012; Kuehner & Brow, 2006; Lu & Lin, 2019; Malabat, Feuerbach, Ma, Saveanu, & Jacquier, 2015; McNeil & Smith, 1985; Policastro, Raborn, Brendel, & Zentner, 2020; Chenxi Qiu et al., 2020; Smale & Baltimore, 1989; Suzuki et al., 2001; Vvedenskaya, Zhang, et al., 2015; Z. Zhang & Dietrich, 2005; Zheng et al., 2011). Structural work indicates that preference for a $Y_{-1}R_{+1}$ initiator sequence occurs, at least in part, by stacking of a purine NTP bound to template strand position +1 and a purine base at template strand position -1 (Basu et al., 2014; Gleghorn, Davydova, Basu, Rothman-Denes, & Murakami, 2011).

It has long been recognized that DNA sequences at positions other than -1/+1 also contribute to the efficiency of initiation by Pol II. However, the sequence preferences at these positions are not well understood, might extend to higher-order interactions between positions, and likely include species-specific determinants. For example, in *S. cerevisiae*, genome-wide transcriptomic data have revealed a preference for an A:T base pair at position -8, reflected as an A on the coding strand and a T on the template strand (Lu & Lin, 2019; Malabat et al., 2015; Policastro et al., 2020; Chenxi Qiu et al., 2020; Z. Zhang & Dietrich, 2005). Furthermore, promoters with high expression and with initiation focused primarily in a single, efficient TSS tend to show additional A enrichment at positions -7 to -3 on the coding strand (T on the template strand) (Lu & Lin, 2019; Lubliner, Keren, & Segal, 2013; Maicas & Friesen, 1990; Chenxi Qiu et al., 2020; Z. Zhang & Dietrich, 2005). Several factors confound efforts to directly measure the contribution of DNA sequence to the efficiency of initiation by Pol II including promoter chromatin context and TSS position within a promoter. These attributes together might be considered a promoter's architecture. While promoter architectural factors are made especially apparent in yeast--where initiation occurs by promoter scanning—and TSSs are examined in a polar fashion from upstream to downstream, they likely contribute to the efficiency of TSS usage by Pol II in most, if not all, eukaryotes.

An elegant and important study of promoter scanning from Kuehner and Brow established that TSS usage is determined by TSS priority during the scanning process (Kuehner & Brow, 2006). As noted above, scanning proceeds directionally from upstream sequences to those downstream. Therefore, sequences are examined by the transcription machinery in the order in which they are scanned and upstream sequences will have priority over downstream ones, regardless of innate TSS strength. To enable comparison of TSSs with different usages in different positions, Kuehner and Brow introduced a concept of "TSS efficiency," which accounts for how much Pol II reaches a particular TSS in order to determine innate TSS strength (see Figure 1).

Imbalanced promoter sequence distributions imposed by evolutionary processes also limit our ability to determine sequence-activity relationships for initiation. For example, it has been observed that yeast promoters have uneven distribution of bases across promoters and this is most obvious in enrichment for T on the coding strand upstream of the median TSS position and A on the coding strand downstream of the median TSS position at highly expressed promoters, and a paucity of G/C content in general (Dujon, 1996; Lu & Lin, 2020; Lubliner et al., 2013; Maicas & Friesen, 1990; Chenxi Qiu et al., 2020). We previously observed that Pol II activity mutants conferred changes to TSS selection including apparent alterations in usage levels for subsets of TSS motifs. Pol II catalytically hyperactive mutants (termed GOF for "Gain-of-Function") and hypoactive mutants (termed LOF for "Loss-of-Function") showed overall decreased and increased usage for TSSs with an A at position -8 relative to the TSS (-8A TSS), respectively (Chenxi Qiu et al., 2020). However, the biased distribution of bases in yeast promoters leads to a biased distribution of sequence motifs, where preferred TSS motifs show enrichment downstream of the median TSS position and less-preferred motifs show enrichment upstream of the median TSS position (Chenxi Qiu et al., 2020). Because Pol II GOF and LOF mutants were observed to shift TSS usage distributions upstream or

downstream, respectively, we could not determine whether the apparent altered TSS motif preferences were a direct or indirect consequence of Pol II catalytic mutant effects on TSS selection. Moreover, other properties such as biochemistry of scanning processivity (Fazal et al., 2015; Lis, 1993; Zhao et al., 2021), promoter identity (Blazeck, Garg, Reed, & Alper, 2012; Dhillon et al., 2020; Lubliner et al., 2015), or promoter chromatin could also contribute to initiation output. Together, all of these factors would make contribution of primary DNA sequence more difficult to ascertain.

In order to remove contextual differences among promoters we have developed a system to dissect determinants of initiation efficiency within a defined, controlled promoter context. Here, we develop “Pol II MASTER” based on bacterial MASTER (MAssively Systematic Transcript End Readout) (Hochschild, 2015; Vvedenskaya, Goldman, & Nickels, 2018; Vvedenskaya et al., 2016; Vvedenskaya, Zhang, et al., 2015; Winkelman et al., 2016), which allows determination of initiation at base pair resolution and attribution of RNA transcripts to nucleic acid barcoded promoter variants in massively parallel fashion. We apply Pol II MASTER to initiation by promoter scanning to investigate the initiation efficiency of ~ 80,000 promoter variants within an appropriate TSS region in Pol II WT and catalytic mutants. We show that this system enables determination of the interface between initiation factor activity, promoter sequence, and promoter output. We recapitulate the large impact of known base positions relative to the TSS on initiation efficiency (-8, -1/+1) while revealing the wide range of effects other positions have in initiation (-11 to -9, -7 to -2 and +2 to +4). We identify a distinct hierarchy in $Y_{-1}R_{+1}$ preferences, and detect interactions between bases at neighboring positions, suggesting potential mechanistic coupling between positions. We find that Pol II mutant classes increase or decrease initiation efficiencies for all possible sequences, consistent with predictions that the primary effects of altered TSS selection (directional shifts in TSS distributions) are driven by initiation efficiency changes across all sequences and not on TSS sequence selection per se. Our results further show that Pol II activity level does contribute to selective efficiency of initiation at sequences +1A vs +1G. Our findings demonstrate that Pol II MASTER provides a platform for quantitative analysis of how initiation factor activity together with promoter sequence contribute to Pol II transcription initiation *in vivo*.

Results

A high-throughput system for studying TSS sequence effects on Pol II initiation output

TSSs are identified at yeast promoters by a scanning from upstream near the core-promoter to downstream (**Figure 1A**). While promoters are melted starting around +20 from the TATA box (if present), there is a distance restriction that reduces usage of TSSs until they are within a region ~40-140 nt downstream from the core promoter (**Figure 1B**). Downstream DNA is pumped toward the Pol II active site by General Transcription Factor (GTF) TFIIH. As TFIIH remains attached to the PIC, this causes scrunching of DNA within the PIC. One model for this scrunching suggests a large transcription bubble with excess DNA being looped out of the PIC. As the scanning process is controlled biochemically by DNA translocation driven by TFIIH, this means that some TSSs may simply be too far downstream to have a high probability of usage (**Figure 1A**, "unreachable TSS"). Once initiation happens at any site within this window of opportunity for initiation (**Figure 1A, B**), Pol II "flux" – the amount of Pol II progressing to TSSs downstream – is necessarily reduced. Differences between promoter in TSS positioning relative to core promoter proximity restriction, TFIIH processivity, and differences in Pol II flux due to usage at other TSSs mean that DNA sequence effects on initiation are difficult to distinguish. In order to specifically dissect how TSS sequence controls initiation efficiency and how sequence interacts with Pol II catalytic activity in a controlled context, we have established a massively parallel promoter variant assay "Pol II MASTER" where we embed almost all possible sequences within a 9 bp randomized TSS region (**Figure 1C**) constructed on plasmids and introduced into yeast strains with wild type (WT) or mutated Pol II. The sequence libraries constructed are illustrated in **Figure 1C** and are referred to based on their base compositions relative to the coding strand unless otherwise specifically noted. The libraries are referred to specifically based on the identities of bases at positions -8, -1, and +1. The "AYR" library has composition $A_{-8}NNNNNNNY_{-1}R_{+1}$ (N=A, C, G, or T, Y=C or T, R=A or G) relative to the coding strand, with "BYR" having composition $B_{-8}NNNNNNNY_{-1}R_{+1}$ (B=C, G, or T), etc.

This 9 bp randomized TSS region was inserted into a controlled promoter context containing specific functionalities (**Figure 1C**). <u>First</u>, the *GAL1* UAS was utilized to allow expression control of libraries. <u>Second</u>, the TATA-box to TSS region of *SNR37* gene was used as a "scanning region" to direct initiation within the adjacent randomized TSS region. This is because almost no RNA 5' ends are observable from this scanning region within its normal promoter context. <u>Third</u>, the native, highly-efficient *SNR37* TSS region was inserted downstream of the randomized TSS region as a "Flux Detector" (FD). Here, we employ approach of Kuehner and Brow that a highly efficient initiation region placed downstream an upstream TSS region may capture any polymerases that happen to scan past the randomized TSS region. Therefore, TSS efficiency, as defined in (Kuehner & Brow, 2006), can be measured for any TSS of interest as Pol II scanning past the TSS should effectively be specified to initiate (**Figure 1C**). Without this FD region, an absence of downstream initiation would render any upstream TSS as apparently highly efficient as there would be no point of reference. By measuring TSS

efficiency using our "Flux Detector" we can compare TSSs at different upstream positions within promoters and/or across libraries. Fourth, a 24 bp DNA barcode containing 20 positions of randomized bases and 4 interspersed fixed bases (to exclude low-complexity sequences) allows RNA products to be assigned to respective randomized TSS DNA templates. An RNA barcode is critical as bases upstream of the TSS will not be present in the transcribed RNA but are critical for specifying TSS efficiency. Fifth, the *GFP* coding region and *CYC1* terminator have been added to support termination and stabilize RNA products. Libraries were constructed by PCR sewing followed by cloning into plasmid backbone. After amplification in *E. coli*, plasmid libraries were transformed into Pol II WT and mutant yeast strains in triplicate. Library expression was induced by addition of galactose to the medium (4% final) for three hours. Both plasmid DNA and RNA products were extracted from harvested yeast cells and amplified for DNA-seq and TSS-seq (**Figure S1A**).

Several measures indicate high level of reproducibility and coverage depth of library variants (**Figure 1D, Figure S1**). Base coverage in the randomized region was highly even (**Figure S1B**). Correlation analysis for DNA-seq reads per promoter variant suggested that yeast transformation did not alter promoter variant distribution (**Figure S1C, D**). Bulk primer extension of libraries illustrated their average behavior and the amount of initiation deriving from the randomized region (**Figure S1E**). As designed, only a very small fraction of initiation was generated from the barcode region or further downstream, validating that the flux detector captured scanning polymerases (**Figure S1E, F**). Aggregate distribution of reads in our three libraries shows that as TSSs decreased in efficiency from the most efficient library ("AYR") to least ("ARY") reads shift from the designed +1 TSS to downstream positions (**Figure 1D**, left). The apparent shift to upstream position -1 in the ARY library is because purine at designated -1 position serves as the +1 for newly created TSSs. **Figure 1D** (middle) illustrates the aggregate TSS efficiency of each library based on usage at the designated +1 TSS relative to usage at that position plus all downstream usage. As an example of high reproducibility, correlation analysis for efficiency of library "major" TSSs (designed +1 TSS of "AYR" and "BYR" libraries, +2 TSS of "ARY" library) demonstrated that biological replicates were reproducible (**Figure 1D**, right). Therefore, we summed reads from three biological replicates, keeping TSSs that contained at least five TSS-seq reads in each replicate and whose Coefficient of Variation (CV) of TSS-seq reads across replicates was less than 0.5 as a proxy for reproducible behavior (**Figure S1G**). As a result, 97% of possible TSS promoter variants were covered in each library on average (**Table S3**).

Sequence-dependent control of TSS efficiency in *S. cerevisiae*

To ask how our libraries recapitulated known TSS efficiency measurements, we first examined core sequences in our library for the *SNR14* TSS and variants examined by Kuehner and Brow (Kuehner & Brow, 2006) for TSS efficiency (**Figure 2A**). Our randomized library contains the *SNR14* TSS sequence embedded in our *SNR37* context along with all single substitution variants of this sequence, including the subset previously examined in *SNR14.* We found that Pol II MASTER recapitulated the single

base effects on TSS efficiency previously observed while also indicating that single base changes around a TSS can have large effects on TSS efficiency.

Examining all designated +1 TSS variants present in our library, we focused our analysis on positions -8, -1 and +1, which *in vivo* genome-wide data suggested are important determinants for TSS selection (**Figure 2B, C**). We first determined the TSS efficiencies of the designed +1 TSS for all promoter variants, dividing all variants into groups defined by bases at positions -8, -1 and +1 relative to TSS to examine the TSS efficiency distribution within each subgroup. Our results not only recapitulate the importance of these three positions in our defined promoter context but also demonstrate that surrounding positions have a considerable impact on TSS efficiency (note the wide range of efficiencies within each TSS group, **Figure 2B**). First, in our controlled context for TSSs at the designed +1 position, we found that $Y_{-1}R_{+1}$ was essential for initiation above a minimal background relative to $R_{-1}Y_{+1}$. Even in the presence of an A_{-8}, $R_{-1}Y_{+1}$ variants showed essentially no usage. Second, we quantified the very large effect of -8A on TSS efficiency (note that $A_{-8}Y_{-1}R_{+1}$ motifs were much higher in efficiency in aggregate than non-$A_{-8}Y_{-1}R_{+1}$ TSSs), demonstrating that -8A alters TSS efficiency. This result is in line with the observation that $A_{-8}C_{-1}A_{+1}$ motif-containing TSSs have the highest aggregate TSS usage from genomic promoters and appears to be the most efficient (Chenxi Qiu et al., 2020). However, usage is a consequence of efficiency and promoter expression level, and genomic promoter efficiency analyses cannot account for other potential sequence or contributions. Third, among $Y_{-1}R_{+1}$ elements ($C_{-1}A_{+1}$, $C_{-1}G_{+1}$, $T_{-1}A_{+1}$, and $T_{-1}G_{+1}$) we found a clear hierarchy of efficiency that was not apparent from genomic promoter TSS usage or contexts, likely due to promoter sequence skew at these promoters. Previous genomic studies indicated $T_{-1}A_{+1}$ and $T_{-1}G_{+1}$ containing TSSs had higher aggregate usage than $C_{-1}G_{+1}$ containing TSSs (Lu & Lin, 2019; Chenxi Qiu et al., 2020). We found that $C_{-1}G_{+1}$ containing TSSs were more efficient than $T_{-1}A_{+1}$ and $T_{-1}G_{+1}$ containing TSSs, suggesting that lower aggregate usage of $C_{-1}G_{+1}$ containing TSSs at genomic promoters likely partially reflects A/T richness of yeast genome (Dujon, 1996).

As noted above, a wide range of TSS efficiencies were observed within each -8/-1/+1 group, even the most efficient $A_{-8}C_{-1}A_{+1}$ group, suggesting meaningful contribution of other positions to initiation efficiency. To ask whether effects of individual -7 to -2 sequences have similar effects regardless of -8, -1 and +1 identity, we rank ordered all individual TSSs by the efficiency of their $A_{-8}C_{-1}A_{+1}$ version (**Figure 2C**). This rank ordering by $A_{-8}C_{-1}A_{+1}$ was predictive of efficiency ranks for -7 to -2 sequences with different bases at positions -8, -1 and +1. This observation indicates important contributions from positions beyond positions -8, -1 and +1 and that positions might function independently to determine TSS efficiency (examined below).

We set out to determine the contributions to Pol II TSS efficiency of individual bases at each position across our randomized region relative to the designed +1 TSS. To do so, we calculated the TSS efficiencies of examined TSS variants and divided them into subgroups based on bases at individual positions relative to the designed TSS +1 position (**Figure 2D**). Comparison of TSS efficiencies across base subgroups suggested significant individual base effects on TSS efficiency in aggregate at all

examined positions. The usage of other positions in our promoter outside of the designed TSS +1 position (either within the randomized region or outside of it) allowed us to examine contribution of bases in our randomized region to the efficiency of these other TSSs (**Figure S1E, F, Figure 2E**). Our combined libraries allowed us to analyze efficiencies of nearly a million TSS variants present within them (distribution of efficiencies is shown in **Figure 2F**). In order to visualize sequence preferences, we used the median initiation efficiency values of each base subgroup as indicators for preference, using centered median values to calculate "relative efficiency", and illustrated this preference in a sequence logo (**Figure 2G**). Datasets of designed +1 TSSs deriving from "AYR", "BYR", and "ARY" libraries allowed us to nearly comprehensively study preference at positions -8 to +1 (**Figure 2H**, middle and bottom). Additionally, 10-25% of total TSS usage among different libraries deriving from a TSS at +4 (**Figure 2E**, brown TSS arrow and sequences) allowed study of positions -11 to -9 relative to this TSS (**Figure 2H**, top). As noted above, positions -8, -1 and +1 are major determinants for TSS efficiency. Interestingly, position -9 showed a relatively strong effect in our defined promoter context, which was not described in genome-wide analyses. At positions -4 to -2, we observed modulation of initiation efficiency, where in general C and/or G were preferred and T was less-preferred. This overall preference is consistent with an observation where individually mutating Ts at positions -4 to -2 relative to -38 TSS of *ADH1* promoter to a C significantly increased usage of that TSS (Faitar, Brodie, & Ponticelli, 2001). Though preferences at positions -7 to -5 are statistically significant (**Figure 2D**), contributions are much lower than other examined positions. Taken together, these results indicate that positions -9 to -8 and -4 to +1 are two major clusters contributing to TSS efficiency.

Experiments across species and the description of the canonical initiator element (Inr) suggest some sequence contributions from downstream positions relative to the TSS (Arkhipova, 1995; Basu et al., 2014; Deshpande & Patel, 2012; Faitar et al., 2001; Gleghorn et al., 2011; Javahery, Khachi, Lo, Zenzie-Gregory, & Smale, 1994; Yarden, Elfakess, Gazit, & Dikstein, 2009; Zheng et al., 2011). To determine the impact of sequences downstream of the TSS, we examined motif enrichment of the most efficient -8 TSS variants, whose positions +1 to +9 are located in the randomized region (**Figure 2E**, green TSS arrow and sequences). We found an $A(/G)_{+2}G(/C)_{+3}G(/C)_{+4}$ enrichment for the top 10% most efficient TSS, but not for the next 10% most efficient TSS (**Figure S2**). These preferences are consistent with early mutagenesis work in yeast, where A(+2) to C/T, G(+3) to T, or C(+4) to T substitutions decreased the utilization of a particular TSS, but A(+2) to G or T(+5) to C substitutions had minor effects (Faitar et al., 2001). This G(/C) enrichment might be related to higher stability of RNA-DNA hybridization or potential direct interaction with Pol II active site (see Discussion).

Pairwise nucleotide-position dependencies have been observed in 5' splice sites (Carmel, Tal, Vig, & Ast, 2004; Roca et al., 2008; Wong, Kinney, & Krainer, 2018). To investigate potential higher order sequence interactions *i.e.* potential coupling between positions contributing to TSS efficiency, we examined all possible pairwise interactions among positions -11 to +1 (**Figure 3, Figure S3**). "Coupling" would entail a base at one position determining the contribution or effect of a base at another position. We found evidence for coupling at multiple positions with the strongest coupling between -9 and -8

positions. Here, an A at one position suppresses the preference of A at the other position suggesting positional epistasis at each position where an A at one position might diminish the impact of an A at the other (**Figure S3A**). Using this -9/-8 interaction as an example, **Figure 3A** shows how coupling is detected and visualized. We calculated "centered relative efficiency" values for each base and position and visualized as a heatmap (**Figure 3A, B).** We found that evident interactions were mainly observed at neighboring positions (**Figure 3B**), especially within positions -9 to -7 and within positions -5 to -2. In addition to its interaction with position -9 described above, the position -8 was also observed to interact with position -7 (**Figure S3B**). Related to the -8/-7 interaction, a direct contact between template strand bases at positions -7/-8 and GTF TFIIB B-reader residues was observed in a Pol II-TFIIB complex structure (Sainsbury, Niesser, & Cramer, 2013). Together, these observations suggest that positions with observed interactions might work together in TSS selection (see Discussion).

Pol II mutants alter TSS efficiency for all possible TSS motifs while showing selective effects at +1

Pol II mutants were observed previously to change apparent specificity for A_{-8} versus B_{-8} (non-A) TSSs in opposite directions depending on Pol II defect as determined by genomic TSS usage analysis (Chenxi Qiu et al., 2020). As we noted in our prior work, this result could be a consequence of Pol II or GTF mutants shifting TSSs upstream or downstream coupled with uneven distributions of TSS motifs within genomic promoters, and the ability of a -8A to function as an upstream R_{+1} TSS that might lack its own A_{-8}. To determine how Pol II catalytic activity affects TSS selection in our controlled promoter context, we measured effects on TSS efficiency in our promoter variant libraries between WT and Pol II mutants (**Figure 4, Figure S4**). We first investigated the overall effects of Pol II mutants on TSS usage (**Figure S4A**). As observed across the genome, Pol II GOF mutants (G1097D and E1103G) shifted TSS usage upstream in aggregate for all libraries compared to WT, whereas Pol II LOF mutants (F1086S and H1085Q) shifted TSS usage downstream in aggregate across all libraries. We observed high reproducibility across biological replicates (**Figure S4B, C**). Therefore, we aggregated reads from biological replicates and employed cutoffs for total reads and promoters with variance above a threshold determined by the coefficient of variation (**Figure S4D**). Clustering analysis for efficiencies of major TSSs among all strains showed that Pol II WT and mutant classes (GOF and LOF) could be separated into groups based on mutant class (**Figure S4B**), consistent with Pol II mutant classes being distinguishable by a variety of profiling methods (Braberg et al., 2013; C. Qiu et al., 2016; Chenxi Qiu et al., 2020). In total our analysis allowed examination of >900,000 TSS sequences for each Pol II mutant (**Figure 4A**).

We first calculated TSS efficiency changes from WT for each TSS variant and then used the median values of base subgroups at each position to indicate effects (**Figure 4B**). Although slight apparent selective effects for specific sequences could be observed (addressed below), Pol II mutants showed directional effects on TSS efficiencies across all TSS motifs, dependent on predicted changes to Pol II catalytic activity (**Figure 4B, Figure S4E**). Specifically, Pol II hyperactive mutants (G1097D and E1103G) increased

overall efficiency at all TSS sequences, whereas Pol II hypoactive mutants (F1086S and H1085Q) decreased overall efficiency for all TSS sequences. The same direction of effects on preferences for A_{-8} and B_{-8} containing TSSs argues against the hypothesis that observed divergent effects on usage of A_{-8} and B_{-8} containing motifs from genomic promoters were derived from changes to innate preference for TSS motifs. Therefore, the apparent changed selectivity for A_{-8} versus B_{-8} TSSs observed in genomic analysis is likely an indirect effect of Pol II mutants, resulting from polar shifts in TSS distribution relative to the uneven underlying TSS motif distributions. We did observe selective effects of Pol II mutants on TSS efficiencies for specific sequences beyond the uniform polarities of changes observed across all sequences. Specifically, both Pol II GOF and LOF mutants showed apparent reduced effects on A_{-8} compared to non-A_{-8} containing motifs. Additionally, effects on G_{+1} containing motifs compared to A_{+1} motifs appeared reduced as well (**Figure 4B**). Some of these apparent differences could be more apparent than real due to different levels of innate TSS efficiency and range effects. Range effects, such as “ceiling” or “floor” effects, arise from compression in signal range; for example, TSS efficiency cannot be better than 100% or worse than 0%. To visualize this, efficiencies of $N_{-8}Y_{-1}R_{+1}$ variants in mutants were plotted based on their WT efficiency and colored based on -8/-1/+1 subgroups to control for confounding variable of different average efficiencies for particular motifs (**Figure 4C**). We observed two independent and additive patterns related to A_{-8}/B_{-8} and A_{+1}/G_{+1}, discussed below.

First, A_{-8} containing TSSs, including $A_{-8}Y_{-1}A_{+1}$ and $A_{-8}Y_{-1}G_{+1}$, showed higher efficiency than B_{-8} TSSs in aggregate for both Pol II GOF and LOF mutants across all TSS efficiencies. It also appears the relatively smaller Pol II GOF effects on A_{-8} containing TSSs observed in **Figure 4B** are likely due to a “ceiling” on TSS efficiency, where many A_{-8} containing TSSs that are already highly efficient have little room for further increase in efficiency (**Figure 2B, D**). Second, A_{+1} containing TSSs, including $A_{-8}Y_{-1}A_{+1}$ and $B_{-8}Y_{-1}A_{+1}$, showed higher efficiencies than G_{+1} containing TSSs in GOF but lower efficiencies in LOF mutants (**Figure 4B, C**). This provocative result supports potential Pol II active site control over initiation NTP preference where Pol II active site sensing of NTP levels can control initiation decisions in meaningful ways (Jenks, O'Rourke, & Reines, 2008; Kuehner & Brow, 2008). Alternatively, differential usage of +1G versus +1A sites could reflect possible alterations to GTP/ATP ratios in Pol II mutants (see Discussion).

We note that Pol II GOF mutant G1097D, our strongest hyperactive mutant *in vitro* with a severe growth defect *in vivo*, generally increased efficiency to a lesser extent than E1103G (**Figure 4B, Figure S4E, F**). That was unexpected because G1097D showed a higher *in vitro* elongation rate than E1103G and stronger effects on TSS shifts at *ADH1, GAL1*, and *IMD2* (Kaplan et al., 2012) and genome wide (Chenxi Qiu et al., 2020), along with its aforementioned stronger growth defects. Consistent with its usually stronger phenotypes, we observed greater far upstream TSS usage from our promoters in G1097D than E1103G (**Figure S4A**). To reconcile these observations, we speculate that the position of TSS within a promoter or scanning window might also affect TSS efficiency. That means while we observe upstream TSS shifting due to increased overall efficiency there may be complexities in the scanning process such that downstream TSSs are affected differently or appear to have decreased efficiency as initiation approaches background level. Evidence supporting this hypothesis is, at upstream

TSSs, such as designed -32 TSS and -8 TSS, G1097D showed higher efficiency than E1103G (**Figure S4F**). At genomic promoters, GOF mutants were observed to increase TSS efficiency of TSSs upstream of promoter "median" TSSs but decrease efficiency of downstream TSSs relative to WT (Chenxi Qiu et al., 2020).

We did not observe strong effects on sequence preference at positions upstream of TSS or sequence interaction in Pol II mutants (**Figure S4G, H, J, K**), suggesting these attributes do not have specific interactions with altered Pol II activity. However, we did observe different base enrichment at position +2 of the most efficient -8 TSSs in Pol II mutants (**Figure S4I**) compared to that in WT (**Figure S2**), suggesting Pol II active site might interact with this position. Specifically, we observed differences in A/G enrichment at position +2 between WT and Pol II mutants, especially GOF mutants. Compared to WT (**Figure S2**), G1097D decreased +2G enrichment, whereas E1103G increased +2A enrichment. These Pol II mutant alterations on sequence enrichment indicate that Pol II catalytic preferences might be impacted by bases at position +2 and that G1097D and E1103G might alter Pol II activity in allele-specific ways.

Learned initiation preferences are predictive of TSS efficiencies at genomic promoters

To ask how sequence determinants identified here relate to natural promoters, we compared our library-defined sequence efficiencies to TSS efficiencies observed at genomic promoters (**Figure 5**). To limit potentially confounding factors for genomic promoters, we focused on a single "median" TSS for each promoter in a defined set of promoter windows. The median TSS is defined as the promoter position containing the 50th percentile of reads within each promoter window (Chenxi Qiu et al., 2020). We found that Pol II sequence preference at positions around median TSSs was mostly consistent with what we observed in our libraries (**Figure 2H, Figure 5A**). Efficient genomic TSSs appear enriched for A at positions -7 to -5. The A-richness at positions between -10 to -3 and +5 to +10 has been noted in previous studies from our lab (Chenxi Qiu et al., 2020) (**Figure 5B**) and others (Lu & Lin, 2019; Malabat et al., 2015; Z. Zhang & Dietrich, 2005). However, As at positions -7 to -5 appeared neutral in our promoter libraries (**Figure 2H**). The observed A-richness in genome could reflect selection *in vivo* for additional promoter properties, such as providing an easily meltable DNA region, lower nucleosome occupancy or reflect a context dependent role not reflected by our promoter libraries.

We find that the interaction between -9A and -8A discovered in our libraries is also reflected in genomic promoters (**Figure 5C**). We first grouped the median TSSs from the top 20% expressed promoters based on the base at position -9 or -8 and then examined sequence enrichment at the other position. We observed that -9A decreased the enrichment of -8A relative to that when position -9 is not A (**Figure 5C**, left). Moreover, when A was absent from the position -8, much higher enrichment for -9A was observed (**Figure 5C**, right). Together, consistent with -9/-8 interaction discovered in our controlled context, in the genome the presence of an A at either position -9 or -8 suppresses enrichment of A at the other position within the group of most efficiently used median TSSs. These results suggest that -9A may function in similar fashion as -

8A, but that -8A is more efficient and therefore has been evolutionarily favored (see Discussion).

Regression modeling identifies key DNA sequences and interaction for TSS selection regulation

We have found that DNA sequences around the TSS not only additively but also interactively contribute to TSS efficiency. To quantitatively identify key features (sequences and interactions) for TSS efficiency, we compiled datasets deriving from all libraries and predicted TSS efficiency from sequence information by logistic regression coupled with a forward stepwise selection strategy (**Figure 6A-F, Figure S5A, B**). We first compiled datasets generated from designed -8 to +2 and +4 TSSs deriving from "AYR", "BYR", and "ARY" libraries (**Figure 6A**) and split data into training (80%) and test (20%) sets. For each variant, its sequences at positions -11 to +9 were extracted as potential predictors for TSS efficiency. We then used a forward stepwise strategy with a 5-fold Cross-Validation (CV) to select robust features (predictors). By evaluating model performance with R^2, sequences at nine positions (positions -9 to -7 and -4 to +2) and one interaction (-9/-8 interaction) were identified as robust features and selected for final modeling (**Figure 6B**). It is worth noting that models with as few as three features – sequences at positions -9, -8, -1 (or +1) – could explain 74.10% of TSS efficiency variation.

The final model containing the most predictive features explained 91.60% of the variance in TSS efficiency for WT test set (20% of total dataset) (**Figure 6C, Figure S5A**). We next asked whether the features learned by modeling using compiled datasets were consistent with our previous sequence preference analysis using selected representative datasets with the most randomized bases. We centered additive variable values and visualized as a sequence logo (**Figure 6D**). First, as expected, positions -1 and +1 were the major predictors, however the influence of the -8 A did not appear as strong as in our previous preference analysis. We suspected this might be because the -9/-8 interaction contribution was not included. After adding the -9/-8 interaction term, we observed emergence of the position -8 as an influential predictor (**Figure 6E, F**), which also emphasizes the contribution of the -9/-8 interaction. Second, and importantly, modeling confirmed the +2A preference observed in previous motif enrichment analysis using only the most efficient -8 TSS variants (**Figure S2**), and being selected for final model further emphasized importance of position +2. The impact of position +2 is also evident when performing Principal Component Analysis (PCA) to variables of models trained with WT or Pol II mutant datasets (**Figure S5B**). The fact that sequences at position +2 are top contributing variables in 2nd principal component that distinguished G1097D and E1103G is agreement with differentially altered +2 sequence enrichment by Pol II mutants (**Figure S4I**), suggesting position +2 preference is altered by Pol II activity changes, and that this position might work directly with the Pol II active site.

Sequence defines TSS efficiency within a wider promoter context during initiation by scanning

To evaluate extent to which DNA sequence around a TSS contributes to TSS efficiency at genomic promoters, we compared the difference between observed and model predicted efficiencies of all positions within known promoter windows or within specific subgroups of known promoters (**Figure 6G-I, Figure S5C**). As expected, we found most promoter positions showed low or no observed efficiency and were over-predicted by sequence alone (**Figure 6G, Figure S5C**), because TSSs need to be specified by a core promoter and scanning occurs over some distance downstream. Therefore, an individual potential TSS has additional attributes that may affect its efficiency beyond local sequence, such as distance from the site of PIC assembly. We therefore extracted only median TSSs purportedly within promoter scanning windows to ask how our sequence-based predictor functions on genomic TSSs (**Figure 6H**). We also separated median TSSs by promoter classes based on Taf1 enrichment (Rhee & Pugh, 2012), a proxy for the two main types of promoters in yeast, or promoter expression levels (**Figure 6I**). We observed good prediction performance at a wide range of TSSs indicating sequence determinants identified in our limited promoter context contribute to TSS efficiency in genomic promoter contexts. We observed increased performance for higher expressed promoters (Pearson r increased from 0.37-0.54 to 0.46-0.64) and Taf1-depleted promoters (Pearson r increased from 0.37-0.50 to 0.54-0.64) (**Figure 6I**). Conceivably, highly expressed promoters may have evolved TSSs at optimal distances from core promoters, and therefore are may be similarly sensitive to sequence effects. Alternatively, the use in our libraries of *GAL1* UAS and *SNR37* core promoter elements, where both *GAL1* and *SNR37* are highly-expressed Taf1-depleted promoters, may share sequence sensitivities for TSSs from related promoters.

Discussion

Individual TSS sequences are critically important for initiation output but this has not been systematically determined for eukaryotic promoters. Changes to initiation factor activity may also alter initiation preferences due to functional if not physical interaction with DNA sequence. Studies examining TSS selection have been based on existing gene promoters *in vivo* and are subject to unknown biases or confounding variables of different promoter contexts. Here we developed and employed Pol II MASTER to systematically investigate ~1 million TSS sequences in wildtype or Pol II mutant cells. This system allowed us to specifically and comprehensively study TSS efficiencies in initiation by promoter scanning, without confounding effects from other architectural features, such as variability in core promoter-TSS distances, differences in promoter identities or chromatin configurations that may obfuscate analyses of genomic TSSs. We find sequence variation at different positions around TSS considerably tunes initiation efficiency in a predictable way and these contributions are important for initiation efficiency at genomic promoters.

Combining results from this study and others, we suggest how TSS sequence contribution for TSS selection works (**Figure 6**). We propose that two major groups contribute to TSS selection: bases around TSS (actual initiating site), bases around position -8. First, in promoter scanning, the TSS and adjacent bases interact with Pol II active site, the 1st NTP or each other to facilitate stable binding of 1st NTP and potentially the 2nd NTP to stimulate RNA synthesis. This would be in contrast to the concept of the initiator or downstream elements functioning as part of the TFIID or PIC binding site as has been proposed for higher eukaryotes (see (Luse, Parida, Spector, Nilson, & Price, 2020; Vo ngoc, Huang, Cassidy, Medrano, & Kadonaga, 2020) and references therein). As the universal initiating element, $Y_{-1}R_{+1}$ has been established to facilitate stable binding of the NTPs by RNA polymerases via base stacking between R_{-1} from template DNA and the 1st purine NTP (Basu et al., 2014; Gleghorn et al., 2011). Positions upstream of the TSS, such as positions -4 to -2, might contribute to stabilize template DNA via base stacking or physical interaction with initiation factors (Basu et al., 2014; Y. Zhang et al., 2012).

Our observation that Pol II mutants showed selective effects on base at position +1 (**Figure 4**) supports that position +1 functions through the Pol II active site and could suggests a mechanism for cellular state to regulate initiation via alteration of initiating base (ATP/GTP) ratios. We have observed that A_{+1} containing TSSs increase in efficiency more than G_{+1} containing TSSs in Pol II GOF mutants, while A_{+1} containing TSSs decrease in efficiency more than G_{+1} containing TSSs in Pol II LOF mutants. In other words, in both classes of Pol II mutants G_{+1} containing TSSs appeared relatively buffered compared to A_{+1} containing TSSs. Two possible explanations could be envisioned. The first explanation is that differential preference for ATP versus GTP is directly affected by altered Pol II activity. The second is that differential effects on TSS efficiency for +1 A sites versus +1 G sites result from indirect effects on ATP/GTP ratio in cells. Such defects might result from altered synthesis of nucleotide synthesis-related genes, a number of which are themselves sensitive to Pol II activity (Braberg et al., 2013; Kaplan et al., 2012; Kuehner & Brow, 2008; Kwapisz et al., 2008; Malik, Qiu,

Snavely, & Kaplan, 2017; Thiebaut et al., 2008). More specifically, a hypothetical decreased GTP/ATP ratio in GOF mutants would suppress increased initiation efficiency of G_{+1} containing TSSs relative to A_{+1} containing TSSs, while a hypothetical increased GTP/ATP ratio in LOF mutants would compensate for a decrease in efficiency of G_{+1} containing TSSs. This explanation could be consistent by multiple observations of ATP- and GTP-related genes in Pol II mutants. In GOF mutants the expression of GTP-related genes, *IMD2*, *IMD3* and *IMD4*, is known to be defective (Braberg et al., 2013; Kaplan et al., 2012), while in Pol II LOF mutants, *IMD2* is constitutively active (Kaplan et al., 2012; Malik et al., 2017) and overall expression at the mRNA level for *IMD* genes is increased (Kaplan et al., 2012). These are consistent with differential effects on GTP relative to ATP synthesis as the products of the *IMD* genes (inosine monophosphate dehydrogenase) compete with the *ADE12* product (adenylosuccinate synthase) for inosine monophosphate (IMP) in the synthesis of GMP and AMP precursors, respectively (Kanehisa & Goto, 2000; Rolfes, 2006; Strathern, Jones, & Broach, 1982). Furthermore, Pol II GOF E1103G was determined to have reduced levels of guanine, which could have knock-on effects to GTP levels or reflect increased demand for guanine-related metabolites (Gout et al., 2017).

The positions +2 to +4 downstream of TSS might contribute to establish NTP stability in Pol II active site as well (**Figure S2**). In addition, we observed Pol II mutant effects on those preferences (**Figure S4I, Figure S5B**), suggesting these positions might function via directly interacting with the Pol II active site, as observed in transcription structures in other species (Basu et al., 2014; Gleghorn et al., 2011). First, T7-like single-subunit RNAP family showed a base-specific interaction between +2 NTP and a residue in the middle of the O helix, which was suggested to enhance formation of the first phosphodiester bond (Gleghorn et al., 2011). Second, a structure of *de novo* transcription initiation complex in bacterial RNA Polymerase showed multiple interactions between the 2nd NTP and its β' and β subunits, whose eukaryotic counterparts are Rpb1 and Rpb2 (Basu et al., 2014). Alternatively, the A/G-rich $A(/G)_{+1}A(/G)_{+2}G_{+3}G_{+4}$ motif might be related to translocation. A study from the Landick lab showed that when A/G comprised the RNA 3' end, the RNAP active site favored the post-translocated state (Hein, Palangat, & Landick, 2011). If A/G similarly effects translocation state within first four bases in Pol II initiation, synthesis of the first few bases might be promoted. Together, positions downstream of TSS might contribute to stabilize or facility first few NTPs adding.

Where TFIIB has been specifically implicated is in bases near position -8, where the TFIIB B-reader domain has been observed to directly interact in a structure of a yeast Pol II-TFIIB complex (Sainsbury et al., 2013). Here, it is attractive to envision TFIIB functioning as an anchor point to pause the scanning process to promote Pol II initiation at a fixed distance downstream. Several observations support this proposed function. First, we detected sequence interaction between positions -8 and -7 (**Figure 3B, Figure S3B**). This is in line with the direct contact of -7T and -8T on the template strand and TFIIB B-reader R64 and D69 observed in Pol II-TFIIB complex structure (Sainsbury et al., 2013), which has been proposed to hold TSSs in Pol II active site longer during scanning. Second, we observed a strong interaction between positions -9 and -8, where the presence of an A at either position suppressed the preference of A at the other

position (**Figure 3, Figure S3A**). This -9/-8 interaction was also evident when examining genomic median TSSs (**Figure 5C**). Taken together, we speculate that Ts around position -8 or -9 on the template strand and TFIIB may pause the scanning process to facilitate the usage of TSSs positioned 8 to 9 bases downstream. Moreover, we have shown that Pol II catalytic mutants alter TSS efficiencies across all TSS sequences, without showing alteration in preference for -8A (**Figure 4B, C, Figure S4E**), suggesting that Pol II catalytic activity is not responsible for -8 preference.

Whether or how DNA sequence surrounding the TSSs is involved in other promoter properties is another question. As at positions -7 to -5 were measured as neutral in our promoter libraries (**Figure 2H**), in contrast to the A-enrichment at highly expressed and focused genomic TSSs (**Figure 5A, B**) (Lu & Lin, 2019; Maicas & Friesen, 1990; Malabat et al., 2015; Chenxi Qiu et al., 2020; Z. Zhang & Dietrich, 2005). We speculate that observed A-richness around TSS functions through other evolved promoter properties. First, observed A-richness between positions -10 to -3, together with T-richness at further upstream core promoter region (Lubliner et al., 2013; Maicas & Friesen, 1990), provides an easily meltable region for DNA unwinding, perhaps facilitating transcription initiation in specific contexts. Second, higher A/T content may function to lower nucleosome occupancy (Segal & Widom, 2009), because appropriately periodic G/C dinucleotides promote nucleosome occupancy (Lee et al., 2007; Peckham et al., 2007; Segal et al., 2006; Tillo & Hughes, 2009). However, the base composition switch in highly expressed promoters from T- to A-preponderance (Maicas & Friesen, 1990; Chenxi Qiu et al., 2020; Wu & Li, 2010) indicates A-richness may have other roles depending on the characteristics of the sequence itself. Third, A-richness may be left over from the evolution of promoter scanning. A recent study of transcription initiation mechanism investigated 12 yeast species and proposed that during evolution a A-rich region upstream of TSS appeared first, then the specific -8A preference occurred (Lu & Lin, 2020). Therefore, the observed A preference at upstream positions of highly used and focused TSSs may be leftovers of A-enrichment in those promoters during scanning evolution in addition to promoter roles beyond TSS selection per se.

Our studies highlight the strength of approaches to minimize contextual factors by isolating specific promoter attributes for study in high-throughput. Here we have employed Pol II MASTER to the DNA sequence determinants of initiation efficiency during Pol II scanning. It will be valuable to apply this systematic analysis to other promoter architectural factors determining Pol II initiation, such as UAS identity, core promoter-TSS distance and sequence composition in scanning region. In addition, we have found sequences downstream of TSSs contribute to TSS efficiency. Therefore, expanding the randomized region is needed to refine our understanding of sequence preference and potential longer range sequence interactions. Furthermore, applying Pol II MASTER across initiation mutants and promoter variants will reveal factor-sequence relationships and may allow initiation potential to be determined from DNA sequence and genome location alone.

Methods

Yeast strains, plasmids, oligonucleotides and media

Yeast strains, plasmids and oligonucleotide sequences are described in **Table S1**. All oligonucleotides were obtained from IDT. Yeast strains used in this study were constructed as previously (Braberg et al., 2013; Jin & Kaplan, 2014; Kaplan et al., 2012; Chenxi Qiu et al., 2020). Briefly, plasmids containing *rpb1* mutants (G1097D, E1103G, F1086S, and H1085Q) were introduced by transformation into yeast strain CKY749 containing a chromosomal deletion of *RPO21/RPB1* but with a wild type *RPB1 URA3* plasmid, which was subsequently lost by plasmid shuffling. Yeast media are following standard protocols (Amberg, Burke, Strathern, Burke, & Cold Spring Harbor Laboratory., 2005). YPD solid medium is made of yeast extract (1% w/v; BD), peptone (2% w/v; BD, 211677), bacto-agar (2% w/v; BD, 214010) and dextrose (2% w/v; VWR, VWRBK876) supplemented with adenine (0.15 mM; Sigma-Aldrich, A9126) and L-tryptophan (0.4 mM; Sigma-Aldrich T0254). Minimal media plates are synthetic complete ("SC") with amino-acids dropped out as appropriate as described in (Amberg et al., 2005) with minor alterations as described in (Kaplan et al., 2012): per standard batch formulation, adenine hemisulfate (Sigma-Aldrich, A9126) was 2 g, L-Leucine (Sigma-Aldrich, L8000) was 4 g, myo-inositol was 0.1 g, para-aminobenzoic acid (PABA) was 0.2 g.

Construction and transformation of plasmid libraries

A 9 nt randomized TSS region and 20 nt randomized barcodes, with 4 fixed bases inserted between every 4 nt (NNNNANNNNCNNNNTNNNNGNNNN), were separately synthesized by IDT as oligo pools with specific randomized positions using "hand mixing" for N positions to ensure even randomization and avoid bias during machine mixing of precursors during oligo synthesis. Together with other components including the *GAL1* UAS, *SNR37* core promoter, *SNR37* TSS region ("flux detector"), *GFP* ORF, and the *CYC1* terminator, template libraries were constructed by PCR sewing and cloned into pRSII413 (a gift from Steven Haase, Addgene plasmid #35450; http://n2t.net/addgene:35450; RRID:Addgene_35450) (Chee & Haase, 2012) by ligation (**Figure S1A**). Ligation products were transformed into *Escherichia coli* TOP10F' cells (Invitrogen) and grown on LB plates supplemented with carbenicillin (100 µg/ml) at high density. 200,000-500,000 colonies were collected from each library to maximize variant representation. Plasmid libraries were isolated from cells pellets using ZymoPURE II Plasmid Maxiprep Kit (Zymo Research, D4203) per manufacturer's instructions. Plasmid library pools were transformed into yeast strains with wildtype and mutated Pol II using chemical transformation and electroporation, respectively. For Pol II WT libraries, 500 ng plasmid pool per reaction was transformed following yeast high efficiency transformation protocol described in (Gietz & Schiestl, 2007). For Pol II mutant libraries, 2 ug plasmid pool per reaction was electroporated into Pol II mutant strains following yeast electroporation transformation protocol described in (Benatuil, Perez, Belk, & Hsieh, 2010), with 50 µg single-stranded carrier DNA added. Transformants were grown on selective SC-His plates with 2% glucose as carbon source at high density. Three biological replicates were performed for each library and on average over two million

colonies were collected for each replicate. Transformants scraped from densely plated transformation plates were inoculated into fresh SC-His medium with 2% raffinose (Amresco, J392) at 0.25 x 10^7 cells/ml and grown until 0.5-0.8 x 10^7 cells/ml, as determined by cell counting. Subsequently, galactose (Amresco, 0637) was added for three hours (4% final concentration) to induce library expression. 50 ml and 5 ml culture aliquots, for RNA and DNA extraction respectively, were harvested and then cell pellets were stored at -80 °C for further processing as described below.

Generation of DNA amplicon for DNA-seq

Plasmid DNA from yeast cell pellets was isolated using YeaStar Genomic DNA Kit (Zymo Research, D2002) per manufacturer's instructions. Amplicon pools containing the TSS and barcode regions were generated using plasmid DNA from *E.coli* or yeast by Micellula DNA Emulsion & Purification (ePCR) Kit (EURx/CHIMERx, 3600) per manufacturer's instructions. To minimize amplification bias, each sample was amplified in a 15-cycle ePCR reaction, purified and subject to an additional 10-cycle scale-up ePCR reaction. To create the necessary sequence template diversity for Illumina sequencing, 18-25 bp and 1-7 bp "stuffer" sequences were added to 5'- and 3'-ends, respectively, during amplicon preparation. Amplicon pools were subject to Illumina NovaSeq 6000 (150 PE) sequencing, and on average 20 M paired-end reads were obtained from each replicate of a sample, with high reproducibility and minimal perturbation of the variant distribution with each library (**Table S2**).

Sample preparation for TSS-seq

Total RNA was extracted by a phenol-chloroform method (Schmitt, Brown, & Trumpower, 1990), followed by RNA purification (RNeasy Mini kit, QIAGEN, 74104) with on-column DNase digestion (RNase-Free DNase Set, QIAGEN, 79254) to remove DNA. TSS-seq was done using procedures described in (Vvedenskaya, Goldman, & Nickels, 2015). To prepare RNAs for the cDNA library construction, samples were sequentially treated with Terminator 5'-Phosphate-Dependent Exonuclease (Lucigen), Quick CIP (calf-intestine alkaline phosphatase, NEB) and Cap-Clip™ Acid Pyrophosphatase (CellScript) to remove 5' monophosphate RNA and convert 5' triphosphate or capped RNAs to 5' monophosphate RNAs. Next, RNA prepared with enzymatic treatments was ligated to the 5'-adapter (s1206-N15, 5'-GUUCAGAGUUCUACAGUCCGACGAUCNNNNNNNNNNNNNNN-3') that contains Illumina adapter sequence and a 15 nt randomized 3'-end to reduce ligation bias and serve as a Unique Molecular Identifier (UMI). Next, cDNA was constructed by reverse transcription using RT primer CKO2191-s128A (5'-CCTTGGCACCCGAGAATTCCAAGTGAATAATTCTTCACCTTTA-3') followed by emulsion PCR amplification for 20-22 cycles using Illumina PCR primers (RP1 and RPI3-30). Final DNA was gel size selected for 180-250 bp lengths and sequenced by Illumina NextSeq 500 (150 SE) or NovaSeq 6000 (200 SE) using custom primer s1115 (5'-CTACACGTTCAGAGTTCTACAGTCCGACGATC-3') to avoid potentially confounding effects of misannealing of the default pooled Illumina sequencing primers to the two randomized sequence regions.

Primer extension assay

Primer extension assays were performed on the same batch of total RNA extracted for TSS-seq as described in (Ranish & Hahn, 1991) with modifications described in (Kaplan et al., 2012). For each reaction, 30 µg total RNA was used. An RNA sample without library transformed was used as “no GFP” control. A sample containing same amount of nuclease-free water was used as “no RNA” control. A primer (CKO2191) complementary to the 6th to 27th bases of *GFP* ORF, which is the same annealing region for reverse transcription of TSS-seq sample preparation, was labeled with ^{32}P γ-ATP (PerkinElmer, BLU502Z250UC) and T4 polynucleotide kinase (Thermo Scientific, EK0031). M-MuLV Reverse Transcriptase (NEB, M0253L), RNase inhibitor (NEB, M0307L), dNTPs (GE) and DTT were added to mix of RNA and labelled primer for reverse transcription reaction. Before loading to sequencing gel, RNase A (Thermo Scientific, EN0531) was added to remove RNA. The products were analyzed by 8% acrylamide/bis-acrylamide (19:1 ratio, Bio-Rad, 1610145) gel containing 1x TBE and 7M Urea. Primer extension gel was visualized by Molecular Imager PharosFX™ Plus System (Bio-Rad) and quantified by Image Lab (5.2).

Computational analyses

Data and statistical analyses were performed in Python (3.8.5) and R (4.0.0) environments. Additional packages usages are reported throughout the methods description. Source code is provided at https://github.com/Kaplan-Lab-Pitt/PolII_MASTER-TSS_sequence. Raw sequencing data and processed data have been deposited on the NCBI SRA (Sequence Read Archive) and GEO (Gene Expression Omnibus) database, under the BioProject accession number PRJNA766624. Visualizations were compiled in Adobe Illustrator 2021.

DNA-seq analysis. High-throughput sequencing of template DNA amplicon was used to assign each 9 nt randomized TSS sequence to a corresponding 24 nt barcode. First, paired-end reads were merged using PEAR (0.9.11) (J. Zhang, Kobert, Flouri, & Stamatakis, 2014). Next, we considered only those reads that contained a perfect match to three sequence regions common to all variants: 27 nt sequence upstream of the TSS region, 24 nt sequence between TSS region and barcode, and 27 nt sequence downstream of barcode (5’-TTCAAATTTTTCTTTTGATTTTTTTTCNNNNNNNNNACATTTTCAAAAGGCTAACATCAGNNNNANNNNCNNNNTNNNNGNNNNATGTCTAAAGGTGAAGAATTATTCACT-3’, randomized TSS and barcode regions are underlined). From these reads, 9 nt TSS region and 24 nt barcode were extracted, followed by individually error correction using UMI-tools (1.0.0) (Smith, Heger, & Sudbery, 2017). Next, for barcodes linking to multiple TSS variants, only barcodes for which >= 90% of the sequencing reads containing a specified barcode also contained a shared, exact 9 nt TSS region were kept. To generate a master pool of TSS-barcode linkages for all TSS-seq samples, for each library (“AYR”, “BYR”, “ARY”), TSS-barcode linkages that existed in at least two out of four samples (one *E.coli* sample plus three WT yeast replicates) and in which >= 5 reads existed were kept and pooled. Two types of processed data are available in GEO database, with accession numbers listed in (**Table S2**): tables containing TSS-barcode

linkages and corresponding DNA-seq read counts for each sample, tables of the master pool containing kept TSS-barcode linkages and corresponding DNA-seq read count in all related samples.

TSS-seq analysis for libraries. High-throughput sequencing of RNA samples was used to link RNA products to barcodes, therefore assign TSS usage to corresponding DNA templates. We considered only those reads that contained a perfect match to a 27 nt sequence region downstream of the barcode, as well as expected length of 5'-end: 5'-[15 nt 5'-UMI]-[>1 nt upstream of barcode region, designated as "RNA 5'-end"]-[24 nt barcode]-[the first 27 nt of GFP ORF, ATGTCTAAAGGTGAAGAATTATTCACT]-3'. Next, 15 nt 5'-UMIs, "RNA 5'-end" with varying length, and 24 nt barcode were extracted and individually corrected by UMI-tools (1.0.0). Deduplication was performed based on 5'-UMIs, meaning reads contained a shared UMI-"RNA 5'-end"-barcode linkage were counted as one deduplicated read for further analysis. Next, the identity of the 24 nt barcode was used to determine the template sequences of randomized TSS region. Then, reads with "RNA 5'-end" sequence perfectly matched to corresponding template sequence were kept and used for downstream analysis. A TSS-seq count table containing TSS usage distribution of each TSS promoter variant was generated. In the count table, each row represents one TSS promoter variant, and each column represents one position between positions -68 to +25 relative to "designed" +1 TSS. The number in each cell represents TSS-seq reads generated from a particular position, with perfectly match to DNA template. After investigating reproducibility, count tables generated from three biological replicates were merged into one by aggregating read counts at each position. Promoter variants with >= 5 TSS-seq reads in each replicate and whose Coefficient of Variation (CV) of TSS-seq reads is less than 0.5 were kept. Two types of processed data are available in GEO database, with accession numbers listed in (**Table S2**): tables containing "RNA 5'-end"-Barcode linkages and corresponding deduplicated TSS-seq read counts for individual sample, TSS-seq count tables for individual samples and after aggregating replicates.

TSS efficiency for each position was calculated by dividing reads count at a particular position by the reads at or downstream of this TSS. TSS positions with >= 20% efficiency but with <= 5 reads left for this and following positions were filtered out, as well as their downstream positions.

Sequence preference analysis. For sequence preference at each position, all TSS variants were subgrouped based on the bases at a particular position. TSS efficiencies of TSS variants were visualized as scatter plots using GraphPad Prism 9. Kruskal-Wallis with Dunn's test was performed to test sequence preference in GraphPad Prism 9. Next, TSS efficiency medians of each subgroup were calculated and centered to calculate "relative efficiency" at each position. The relative efficiencies were visualized as sequence logos using Logomaker (0.8) (Tareen & Kinney, 2020). In motif enrichment analysis, surrounding sequences relative to examined TSSs were extracted and visualized as sequence logos using WebLogo 3 (Crooks, Hon, Chandonia, & Brenner, 2004). Heatmaps, scatter plots and density plots for comparing Pol II WT and mutants were generated by Morpheus (https://software.broadinstitute.org/morpheus) or ggplot2 (3.3.3) R package.

***Interaction analysis*.** The interaction between positions is defined as different bases existing at one position resulting in different sequence preferences at another position. For any two positions, all TSS variants were subgrouped based on bases at both positions. Median values of TSS efficiency distribution of each subgroup were calculated and centered twice to calculate “centered relative efficiency”. The centered relative efficiencies were visualized as heatmaps using Seaborn (0.11.0). Interactions related to positions between -11 to -9 were calculated using datasets of designed +4 TSS deriving from “AYR”, “BYR” and “ARY” libraries. Other interactions were calculated using datasets of designed +1 TSS deriving from “AYR” and “BYR” libraries.

TSS-seq analysis for genomic TSSs. Genomic TSS-seq datasets are from our lab’s previous study (Zhao et al., 2021). Quality control, read trimming and mapping were performed as described in (Chenxi Qiu et al., 2020; Zhao et al., 2021) to generate a TSS count table that contains TSS-seq reads at each individual position within known promoter windows (“median” TSS, 250 nt upstream and 150 nt downstream from median TSS position). TSS efficiency calculation and subsequent sequence preference analyses were performed as that for Pol II MASTER libraries.

Prediction of TSS efficiency. To prepare datasets for modeling, positions of designed -8 to +2 and +4 TSSs of each promoter variant that have valid TSS efficiency were compiled as sequence variants. For each TSS variant, sequences at -11 to +9 positions relative to TSS, together with corresponding TSS efficiency, were extracted. 80% of dataset were randomly partitioned as training set and the rest 20% as testing set. To select robust features, a forward stepwise strategy with a 5-fold Cross-Validation (CV) was employed in two major stages, for additive terms and for interactions. Starting with no variable in the model, logistic regression models with one additional variable (the sequence at a particular position) were trained to predict TSS efficiency on training set by train() of caret (6.0.86) R package (Kuhn, 2008), with a 5-fold CV. The R^2, representing the proportion of variance explained, was calculated to indicate the performance of each model. The variable that provides the highest increased R^2 for model was added into the model for next round of variable selection. This process was repeated until the increased R^2 is less than 0.01. After identifying the most influential additive variables, same process was repeated for investigating robust interactions between selected additive variables. Next, a final model with selected robust features. including additive variables and interactions, was constructed on entire training set using glm() and investigated on testing set. Comparison between predicted and measured efficiencies was visualized as scatter plots using LSD (4.1.0) R package. Model parameters were extracted and used to further calculation. Visualizations were done in Logomaker (0.8) and Seaborn (0.11.0) in Python. Principal Component Analysis (PCA) was performed using prcomp() in R.

Supplementary Tables

Supplementary Table 1. Oligonucleotides, yeast strains, and plasmids.

Supplementary Table 2. Statistics and data accession information for DNA-seq and TSS-seq.

Supplementary Table 3. Summary of libraries.

Acknowledgements

We thank Kaplan and Nickels lab members for discussions and comments on the manuscript.

Authors' contributions

Y.Z. designed the project, performed experiments, analyzed data, made figures, drafted and revised the manuscript. I.V. generated libraries for TSS-seq. B.E.N. provided funding and methodology of TSS-seq, contributed to revise the manuscript. C.D.K. conceived and designed the project, guided analyses and interpretation of data, provided funding, revised the manuscript.

Funding

We acknowledge support from NIH grant R01GM097260 to C.D.K. for the early part of this work and NIH grants R01GM120450 to C.D.K. and R35GM118059 to B.E.N.

Availability of data and materials

Raw sequencing data and reported processed data are available under the in the NCBI BioProject, under the accession number PRJNA766624. Source code for analysis is provided at https://github.com/Kaplan-Lab-Pitt/PolII_MASTER-TSS_sequence.

References

Amberg, D. C., Burke, D., Strathern, J. N., Burke, D., & Cold Spring Harbor Laboratory. (2005). *Methods in yeast genetics : a Cold Spring Harbor Laboratory course manual* (2005 ed.). Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Arkhipova, I. R. (1995). Promoter elements in Drosophila melanogaster revealed by sequence analysis. *Genetics, 139*(3), 1359-1369. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7768444

Basu, R. S., Warner, B. A., Molodtsov, V., Pupov, D., Esyunina, D., Fernandez-Tornero, C., . . . Murakami, K. S. (2014). Structural basis of transcription initiation by bacterial RNA polymerase holoenzyme. *J Biol Chem, 289*(35), 24549-24559. doi:10.1074/jbc.M114.584037

Batut, P., Dobin, A., Plessy, C., Carninci, P., & Gingeras, T. R. (2013). High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res, 23*(1), 169-180. doi:10.1101/gr.139618.112

Benatuil, L., Perez, J. M., Belk, J., & Hsieh, C. M. (2010). An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng Des Sel, 23*(4), 155-159. doi:10.1093/protein/gzq002

Blazeck, J., Garg, R., Reed, B., & Alper, H. S. (2012). Controlling promoter strength and regulation in Saccharomyces cerevisiae using synthetic hybrid promoters. *Biotechnol Bioeng, 109*(11), 2884-2895. doi:10.1002/bit.24552

Boyd, M., Thodberg, M., Vitezic, M., Bornholdt, J., Vitting-Seerup, K., Chen, Y., . . . Sandelin, A. (2018). Characterization of the enhancer and promoter landscape of inflammatory bowel disease from human colon biopsies. *Nat Commun, 9*(1), 1661. doi:10.1038/s41467-018-03766-z

Braberg, H., Jin, H., Moehle, E. A., Chan, Y. A., Wang, S., Shales, M., . . . Krogan, N. J. (2013). From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. *Cell, 154*(4), 775-788. doi:10.1016/j.cell.2013.07.033

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol, 212*(4), 563-578. doi:10.1016/0022-2836(90)90223-9

Carmel, I., Tal, S., Vig, I., & Ast, G. (2004). Comparative analysis detects dependencies among the 5' splice-site positions. *RNA, 10*(5), 828-840. doi:10.1261/rna.5196404

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., . . . Genome Science, G. (2005). The transcriptional landscape of the mammalian genome. *Science, 309*(5740), 1559-1563. doi:10.1126/science.1112014

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., . . . Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet, 38*(6), 626-635. doi:10.1038/ng1789

Chee, M. K., & Haase, S. B. (2012). New and Redesigned pRS Plasmid Shuttle Vectors for Genetic Manipulation of Saccharomycescerevisiae. *G3 (Bethesda), 2*(5), 515-526. doi:10.1534/g3.111.001917

Chen, R. A., Down, T. A., Stempor, P., Chen, Q. B., Egelhofer, T. A., Hillier, L. W., . . . Ahringer, J. (2013). The landscape of RNA polymerase II transcription initiation in

C. elegans reveals promoter and enhancer architectures. *Genome Res, 23*(8), 1339-1347. doi:10.1101/gr.153668.112
Cheng, Z., Otto, G. M., Powers, E. N., Keskin, A., Mertins, P., Carr, S. A., . . . Brar, G. A. (2018). Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis. *Cell, 172*(5), 910-923 e916. doi:10.1016/j.cell.2018.01.035
Chia, M., Li, C., Marques, S., Pelechano, V., Luscombe, N. M., & van Werven, F. J. (2021). High-resolution analysis of cell-state transitions in yeast suggests widespread transcriptional tuning by alternative starts. *Genome Biol, 22*(1), 34. doi:10.1186/s13059-020-02245-3
Consortium, F., the, R. P., Clst, Forrest, A. R., Kawaji, H., Rehli, M., . . . Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature, 507*(7493), 462-470. doi:10.1038/nature13182
Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C., & Chambon, P. (1980). Promoter sequences of eukaryotic protein-coding genes. *Science, 209*(4463), 1406-1414. doi:10.1126/science.6251548
Cortes, T., Schubert, O. T., Rose, G., Arnvig, K. B., Comas, I., Aebersold, R., & Young, D. B. (2013). Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in Mycobacterium tuberculosis. *Cell Rep, 5*(4), 1121-1131. doi:10.1016/j.celrep.2013.10.031
Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res, 14*(6), 1188-1190. doi:10.1101/gr.849004
Demircioglu, D., Cukuroglu, E., Kindermans, M., Nandi, T., Calabrese, C., Fonseca, N. A., . . . Goke, J. (2019). A Pan-cancer Transcriptome Analysis Reveals Pervasive Regulation through Alternative Promoters. *Cell, 178*(6), 1465-1477 e1417. doi:10.1016/j.cell.2019.08.018
Deshpande, A. P., & Patel, S. S. (2012). Mechanism of transcription initiation by the yeast mitochondrial RNA polymerase. *Biochim Biophys Acta, 1819*(9-10), 930-938. doi:10.1016/j.bbagrm.2012.02.003
Dhillon, N., Shelansky, R., Townshend, B., Jain, M., Boeger, H., Endy, D., & Kamakaka, R. (2020). Permutational analysis of Saccharomyces cerevisiae regulatory elements. *Synth Biol (Oxf), 5*(1), ysaa007. doi:10.1093/synbio/ysaa007
Dujon, B. (1996). The yeast genome project: what did we learn? *Trends Genet, 12*(7), 263-270. doi:10.1016/0168-9525(96)10027-5
Faitar, S. L., Brodie, S. A., & Ponticelli, A. S. (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. *Molecular and Cellular Biology, 21*(14), 4427-4440. doi:Doi 10.1128/Mcb.21.14.4427-4440.2001
Fazal, F. M., Meng, C. A., Murakami, K., Kornberg, R. D., & Block, S. M. (2015). Real-time observation of the initiation of RNA polymerase II transcription. *Nature, 525*(7568), 274-277. doi:10.1038/nature14882
Furter-Graves, E. M., & Hall, B. D. (1990). DNA sequence elements required for transcription initiation of the Schizosaccharomyces pombe ADH gene in Saccharomyces cerevisiae. *Mol Gen Genet, 223*(3), 407-416. doi:10.1007/BF00264447

Gietz, R. D., & Schiestl, R. H. (2007). High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nature Protocols, 2*(1), 31-34. doi:10.1038/nprot.2007.13

Gleghorn, M. L., Davydova, E. K., Basu, R., Rothman-Denes, L. B., & Murakami, K. S. (2011). X-ray crystal structures elucidate the nucleotidyl transfer reaction of transcript initiation using two nucleotides. *Proceedings of the National Academy of Sciences of the United States of America, 108*(9), 3566-3571. doi:10.1073/pnas.1016691108

Gout, J. F., Li, W., Fritsch, C., Li, A., Haroon, S., Singh, L., . . . Vermulst, M. (2017). The landscape of transcription errors in eukaryotic cells. *Sci Adv, 3*(10), e1701484. doi:10.1126/sciadv.1701484

Hahn S, H. E., Guarente L. (1985). Each of three "TATA elements" specifies a subset of the transcription initiation sites at the CYC-1 promoter of Saccharomy cescerevisiae. *Proceedings of the National Academy of Sciences, 82*(24), 8562-8566.

Hampsey, M. (1998). Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol Mol Biol Rev, 62*(2), 465-503. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9618449

Hashimoto, S., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., . . . Matsushima, K. (2004). 5'-end SAGE for the analysis of transcriptional start sites. *Nat Biotechnol, 22*(9), 1146-1149. doi:10.1038/nbt998

Healy, A. M., Helser, T. L., & Zitomer, R. S. (1987). Sequences required for transcriptional initiation of the Saccharomyces cerevisiae CYC7 genes. *Mol Cell Biol, 7*(10), 3785-3791. doi:10.1128/mcb.7.10.3785-3791.1987

Hein, P. P., Palangat, M., & Landick, R. (2011). RNA transcript 3'-proximal sequence affects translocation bias of RNA polymerase. *Biochemistry, 50*(32), 7002-7014. doi:10.1021/bi200437q

Hochschild, A. (2015). Mastering Transcription: Multiplexed Analysis of Transcription Start Site Sequences. *Mol Cell, 60*(6), 829-831. doi:10.1016/j.molcel.2015.12.007

Hoskins, R. A., Landolin, J. M., Brown, J. B., Sandler, J. E., Takahashi, H., Lassmann, T., . . . Celniker, S. E. (2011). Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res, 21*(2), 182-192. doi:10.1101/gr.112466.110

Javahery, R., Khachi, A., Lo, K., Zenzie-Gregory, B., & Smale, S. T. (1994). DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol Cell Biol, 14*(1), 116-127. doi:10.1128/mcb.14.1.116-127.1994

Jenks, M. H., O'Rourke, T. W., & Reines, D. (2008). Properties of an intergenic terminator and start site switch that regulate IMD2 transcription in yeast. *Mol Cell Biol, 28*(12), 3883-3893. doi:10.1128/MCB.00380-08

Jin, H., & Kaplan, C. D. (2014). Relationships of RNA polymerase II genetic interactors to transcription start site usage defects and growth in Saccharomyces cerevisiae. *G3 (Bethesda), 5*(1), 21-33. doi:10.1534/g3.114.015180

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28*(1), 27-30. doi:10.1093/nar/28.1.27

Kaplan, C. D., Jin, H., Zhang, I. L., & Belyanin, A. (2012). Dissection of Pol II trigger loop function and Pol II activity-dependent control of start site selection in vivo. *PLoS Genet, 8*(4), e1002627. doi:10.1371/journal.pgen.1002627

Kim, D., Hong, J. S. J., Qiu, Y., Nagarajan, H., Seo, J. H., Cho, B. K., . . . Palsson, B. O. (2012). Comparative Analysis of Regulatory Elements between Escherichia coli and Klebsiella pneumoniae by Genome-Wide Transcription Start Site Profiling. *Plos Genetics, 8*(8). doi:ARTN e1002867

10.1371/journal.pgen.1002867

Kuehner, J. N., & Brow, D. A. (2006). Quantitative analysis of in vivo initiator selection by yeast RNA polymerase II supports a scanning model. *J Biol Chem, 281*(20), 14119-14128. doi:10.1074/jbc.M601937200

Kuehner, J. N., & Brow, D. A. (2008). Regulation of a eukaryotic gene by GTP-dependent start site selection and transcription attenuation. *Mol Cell, 31*(2), 201-211. doi:10.1016/j.molcel.2008.05.018

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, 28*(5), 1-26. doi:DOI 10.18637/jss.v028.i05

Kwapisz, M., Wery, M., Despres, D., Ghavi-Helm, Y., Soutourina, J., Thuriaux, P., & Lacroute, F. (2008). Mutations of RNA polymerase II activate key genes of the nucleoside triphosphate biosynthetic pathways. *EMBO J, 27*(18), 2411-2421. doi:10.1038/emboj.2008.165

Lee, W., Tillo, D., Bray, N., Morse, R. H., Davis, R. W., Hughes, T. R., & Nislow, C. (2007). A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet, 39*(10), 1235-1244. doi:10.1038/ng2117

Lis, C. G. a. J. T. (1993). DNA Melting on Yeast RNA Polymerase II Promoter. *Science, 261*(5122), 759-762. Retrieved from http://www.jstor.org/stable/2882246

Lu, Z., & Lin, Z. (2019). Pervasive and dynamic transcription initiation in Saccharomyces cerevisiae. *Genome Res, 29*(7), 1198-1210. doi:10.1101/gr.245456.118

Lu, Z., & Lin, Z. (2020). The origin and evolution of a distinct mechanism of transcription initiation in yeasts. *Genome Res*. doi:10.1101/gr.264325.120

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*(11), 5569-5581. doi:10.1093/nar/gkt256

Lubliner, S., Regev, I., Lotan-Pompan, M., Edelheit, S., Weinberger, A., & Segal, E. (2015). Core promoter sequence in yeast is a major determinant of expression level. *Genome Research, 25*(7), 1008-1017. doi:10.1101/gr.188193.114

Luse, D. S., Parida, M., Spector, B. M., Nilson, K. A., & Price, D. H. (2020). A unified view of the sequence and functional organization of the human RNA polymerase II promoter. *Nucleic Acids Res, 48*(14), 7767-7785. doi:10.1093/nar/gkaa531

Maicas, E., & Friesen, J. D. (1990). A sequence pattern that occurs at the transcription initiation region of yeast RNA polymerase II promoters. *Nucleic Acids Res, 18*(11), 3387-3393. doi:10.1093/nar/18.11.3387

Malabat, C., Feuerbach, F., Ma, L., Saveanu, C., & Jacquier, A. (2015). Quality control of transcription start site selection by nonsense-mediated-mRNA decay. *eLife, 4*. doi:10.7554/eLife.06722

Malik, I., Qiu, C., Snavely, T., & Kaplan, C. D. (2017). Wide-ranging and unexpected consequences of altered Pol II catalytic activity in vivo. *Nucleic Acids Res, 45*(8), 4431-4451. doi:10.1093/nar/gkx037

McNeil, J. B., & Smith, M. (1985). Saccharomyces cerevisiae CYC1 mRNA 5'-end positioning: analysis by in vitro mutagenesis, using synthetic duplexes with random mismatch base pairs. *Mol Cell Biol, 5*(12), 3545-3551. doi:10.1128/mcb.5.12.3545-3551.1985

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*(7), 603-610. doi:10.1038/nsmb1117

Nepal, C., Hadzhiev, Y., Previti, C., Haberle, V., Li, N., Takahashi, H., . . . Muller, F. (2013). Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. *Genome Res, 23*(11), 1938-1950. doi:10.1101/gr.153692.112

Park, D., Morris, A. R., Battenhouse, A., & Iyer, V. R. (2014). Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. *Nucleic Acids Res, 42*(6), 3736-3749. doi:10.1093/nar/gkt1366

Peckham, H. E., Thurman, R. E., Fu, Y., Stamatoyannopoulos, J. A., Noble, W. S., Struhl, K., & Weng, Z. (2007). Nucleosome positioning signals in genomic DNA. *Genome Res, 17*(8), 1170-1177. doi:10.1101/gr.6101007

Pelechano, V., Wei, W., & Steinmetz, L. M. (2013). Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature, 497*(7447), 127-131. doi:10.1038/nature12121

Policastro, R. A., Raborn, R. T., Brendel, V. P., & Zentner, G. E. (2020). Simple and efficient profiling of transcription initiation and transcript levels with STRIPE-seq. *Genome Res, 30*(6), 910-923. doi:10.1101/gr.261545.120

Qiu, C., Erinne, O. C., Dave, J. M., Cui, P., Jin, H., Muthukrishnan, N., . . . Kaplan, C. D. (2016). High-Resolution Phenotypic Landscape of the RNA Polymerase II Trigger Loop. *PLoS Genet, 12*(11), e1006321. doi:10.1371/journal.pgen.1006321

Qiu, C., Jin, H., Vvedenskaya, I., Llenas, J. A., Zhao, T., Malik, I., . . . Kaplan, C. D. (2020). Universal promoter scanning by Pol II during transcription initiation in Saccharomyces cerevisiae. *Genome Biology, 21*(1). doi:10.1186/s13059-020-02040-0

Ranish, J. A., & Hahn, S. (1991). The Yeast General Transcription Factor Tfiia Is Composed of 2 Polypeptide Subunits. *Journal of Biological Chemistry, 266*(29), 19320-19327. Retrieved from <Go to ISI>://WOS:A1991GK66700034

Rhee, H. S., & Pugh, B. F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature, 483*(7389), 295-301. doi:10.1038/nature10799

Roca, X., Olson, A. J., Rao, A. R., Enerly, E., Kristensen, V. N., Borresen-Dale, A. L., . . . Sachidanandam, R. (2008). Features of 5'-splice-site efficiency derived from disease-causing mutations and comparative genomics. *Genome Res, 18*(1), 77-87. doi:10.1101/gr.6859308

Rojas-Duran, M. F., & Gilbert, W. V. (2012). Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA, 18*(12), 2299-2305. doi:10.1261/rna.035865.112

Rolfes, R. (2006). Regulation of purine nucleotide biosynthesis: in yeast and beyond.

Sainsbury, S., Niesser, J., & Cramer, P. (2013). Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. *Nature, 493*(7432), 437-440. doi:10.1038/nature11715

Schmitt, M. E., Brown, T. A., & Trumpower, B. L. (1990). A rapid and simple method for preparation of RNA from Saccharomyces cerevisiae. *Nucleic Acids Res, 18*(10), 3091-3092. doi:10.1093/nar/18.10.3091

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., . . . Widom, J. (2006). A genomic code for nucleosome positioning. *Nature, 442*(7104), 772-778. doi:10.1038/nature04979

Segal, E., & Widom, J. (2009). Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struct Biol, 19*(1), 65-71. doi:10.1016/j.sbi.2009.01.004

Smale, S. T., & Baltimore, D. (1989). The "initiator" as a transcription control element. *Cell, 57*(1), 103-113. doi:10.1016/0092-8674(89)90176-1

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res, 27*(3), 491-499. doi:10.1101/gr.209601.116

Strathern, J. N., Jones, E. W., & Broach, J. R. (1982). *The Molecular biology of the yeast Saccharomyces : metabolism and gene expression*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory.

Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., . . . Sugano, S. (2001). Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep, 2*(5), 388-393. doi:10.1093/embo-reports/kve085

Tareen, A., & Kinney, J. B. (2020). Logomaker: beautiful sequence logos in Python. *Bioinformatics, 36*(7), 2272-2274. doi:10.1093/bioinformatics/btz921

Thiebaut, M., Colin, J., Neil, H., Jacquier, A., Seraphin, B., Lacroute, F., & Libri, D. (2008). Futile cycle of transcription initiation and termination modulates the response to nucleotide shortage in S. cerevisiae. *Mol Cell, 31*(5), 671-682. doi:10.1016/j.molcel.2008.08.010

Thorsen, K., Schepeler, T., Oster, B., Rasmussen, M. H., Vang, S., Wang, K., . . . Andersen, C. L. (2011). Tumor-specific usage of alternative transcription start sites in colorectal cancer identified by genome-wide exon array analysis. *BMC Genomics, 12*, 505. doi:10.1186/1471-2164-12-505

Tillo, D., & Hughes, T. R. (2009). G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics, 10*, 442. doi:10.1186/1471-2105-10-442

Vo ngoc, L., Huang, C. Y., Cassidy, C. J., Medrano, C., & Kadonaga, J. T. (2020). Identification of the human DPR core promoter element using machine learning. *Nature, 585*(7825), 459-463. doi:10.1038/s41586-020-2689-7

Vvedenskaya, I. O., Goldman, S. R., & Nickels, B. E. (2015). Preparation of cDNA libraries for high-throughput RNA sequencing analysis of RNA 5' ends. *Methods Mol Biol, 1276*, 211-228. doi:10.1007/978-1-4939-2392-2_12

Vvedenskaya, I. O., Goldman, S. R., & Nickels, B. E. (2018). Analysis of Bacterial Transcription by "Massively Systematic Transcript End Readout," MASTER. *Methods Enzymol, 612*, 269-302. doi:10.1016/bs.mie.2018.09.007

Vvedenskaya, I. O., Vahedian-Movahed, H., Zhang, Y., Taylor, D. M., Ebright, R. H., & Nickels, B. E. (2016). Interactions between RNA polymerase and the core recognition element are a determinant of transcription start site selection. *Proc Natl Acad Sci U S A, 113*(21), E2899-2905. doi:10.1073/pnas.1603271113

Vvedenskaya, I. O., Zhang, Y., Goldman, S. R., Valenti, A., Visone, V., Taylor, D. M., . . . Nickels, B. E. (2015). Massively Systematic Transcript End Readout, "MASTER": Transcription Start Site Selection, Transcriptional Slippage, and Transcript Yields. *Mol Cell, 60*(6), 953-965. doi:10.1016/j.molcel.2015.10.029

Winkelman, J. T., Vvedenskaya, I. O., Zhang, Y., Zhang, Y., Bird, J. G., Taylor, D. M., . . . Nickels, B. E. (2016). Multiplexed protein-DNA cross-linking: Scrunching in transcription start site selection. *Science, 351*(6277), 1090-1093. doi:10.1126/science.aad6881

Wong, M. S., Kinney, J. B., & Krainer, A. R. (2018). Quantitative Activity Profile and Context Dependence of All Human 5' Splice Sites. *Mol Cell, 71*(6), 1012-1026 e1013. doi:10.1016/j.molcel.2018.07.033

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*(4), 473-484. doi:10.1101/gr.103226.109

Yamashita, R., Sathira, N. P., Kanai, A., Tanimoto, K., Arauchi, T., Tanaka, Y., . . . Suzuki, Y. (2011). Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res, 21*(5), 775-789. doi:10.1101/gr.110254.110

Yarden, G., Elfakess, R., Gazit, K., & Dikstein, R. (2009). Characterization of sINR, a strict version of the Initiator core promoter element. *Nucleic Acids Res, 37*(13), 4234-4246. doi:10.1093/nar/gkp315

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics, 30*(5), 614-620. doi:10.1093/bioinformatics/btt593

Zhang, P., Dimont, E., Ha, T., Swanson, D. J., Consortium, F., Hide, W., & Goldowitz, D. (2017). Relatively frequent switching of transcription start sites during cerebellar development. *BMC Genomics, 18*(1), 461. doi:10.1186/s12864-017-3834-z

Zhang, Y., Feng, Y., Chatterjee, S., Tuske, S., Ho, M. X., Arnold, E., & Ebright, R. H. (2012). Structural basis of transcription initiation. *Science, 338*(6110), 1076-1080. doi:10.1126/science.1227786

Zhang, Z., & Dietrich, F. S. (2005). Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE. *Nucleic Acids Res, 33*(9), 2838-2851. doi:10.1093/nar/gki583

Zhao, T., Vvedenskaya, I. O., Lai, W. K., Basu, S., Pugh, B. F., Nickels, B. E., & Kaplan, C. D. (2021). Ssl2/TFIIH function in transcription start site scanning by RNA Polymerase II in Saccharomyces cerevisiae. *eLife, 10*. doi:10.7554/eLife.71013

Zheng, H., Sun, L., Peng, W., Shen, Y., Wang, Y., Xu, B., . . . Wang, S. (2011). Global identification of transcription start sites in the genome of Apis mellifera using

5'LongSAGE. *J Exp Zool B Mol Dev Evol, 316*(7), 500-514.
doi:10.1002/jez.b.21421

Figures and Figure Legends

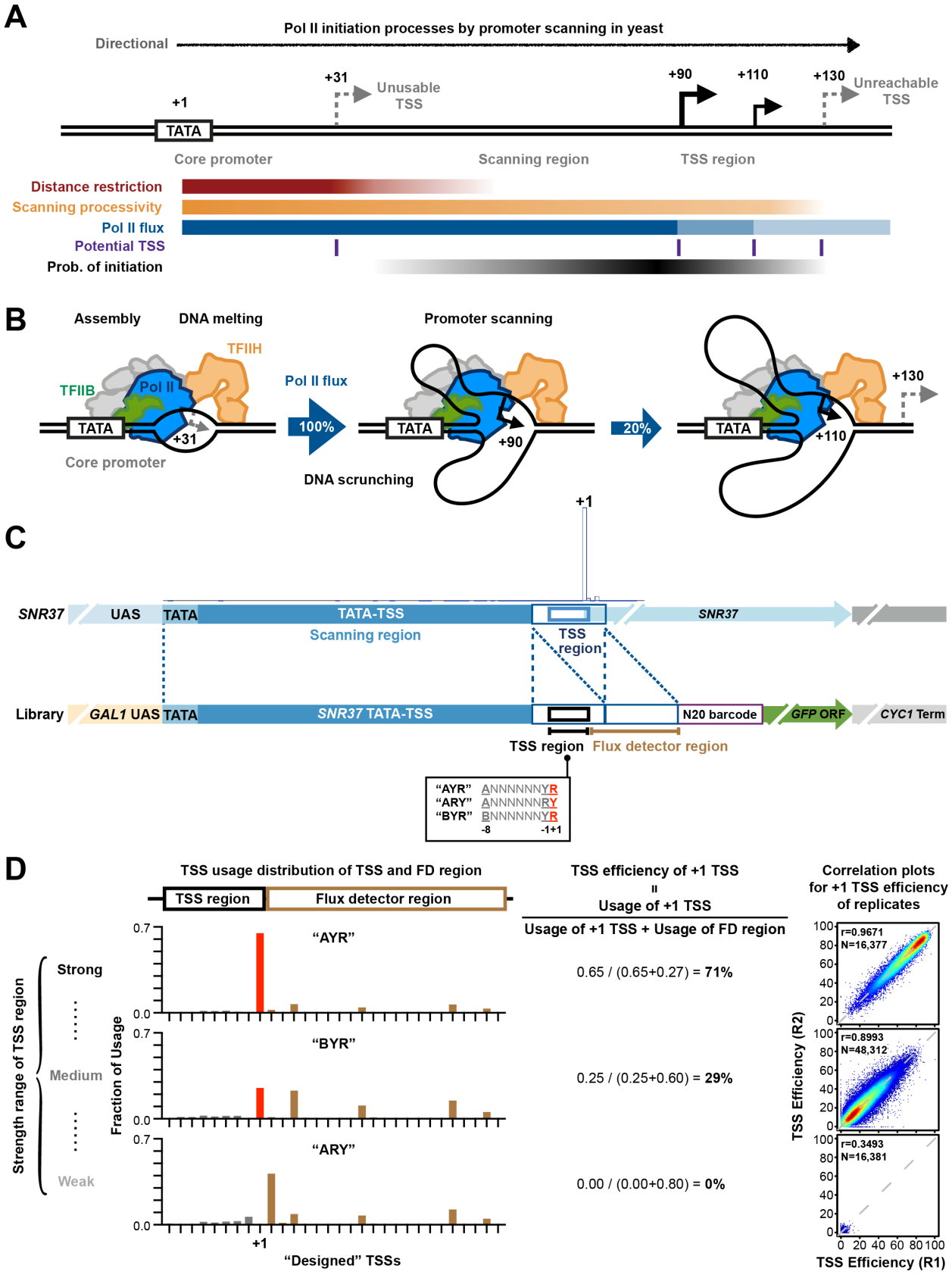

A
Pol II initiation processes by promoter scanning in yeast
Directional
+1
TATA
+31
Unusable TSS
+90
+110
+130
Unreachable TSS
Core promoter
Scanning region
TSS region
Distance restriction
Scanning processivity
Pol II flux
Potential TSS
Prob. of initiation
B
Assembly
DNA melting
Promoter scanning
TFIIH
TFIIB
Pol II
TATA
+31
Core promoter
Pol II flux
100%
DNA scrunching
+90
20%
+110
+130
C
+1
SNR37
UAS
TATA
TATA-TSS
Scanning region
TSS region
SNR37
Library
GAL1 UAS
SNR37 TATA-TSS
N20 barcode
GFP ORF
CYC1 Term
TSS region
Flux detector region
"AYR" ANNNNNNYR
"ARY" ANNNNNNRY
"BYR" BNNNNNNYR
-8
-1+1
D
TSS usage distribution of TSS and FD region
TSS region
Flux detector region
Strength range of TSS region
Strong
Medium
Weak
Fraction of Usage
"AYR"
"BYR"
"ARY"
0.7
0.0
+1
"Designed" TSSs
TSS efficiency of +1 TSS
=
Usage of +1 TSS
Usage of +1 TSS + Usage of FD region
0.65 / (0.65+0.27) = 71%
0.25 / (0.25+0.60) = 29%
0.00 / (0.00+0.80) = 0%
Correlation plots for +1 TSS efficiency of replicates
r=0.9671
N=16,377
r=0.8993
N=48,312
r=0.3493
N=16,381
TSS Efficiency (R2)
TSS Efficiency (R1)
0 20 40 60 80 100

Figure 1. A high-throughput system for studying transcription TSS selection.
(A and B) Pol II initiation in yeast proceeds by a promoter scanning mechanism. Yeast Pol II initiation usually occurs at multiple TSSs ranging from 40 to 120 bp downstream of a core promoter region comprising the PIC assembly position and typically a TATA-box for TATA-containing promoters. After the PIC assembles upstream (core promoter) scanning will proceed over a scanning region towards positions where TSS selection occurs (TSS region). In addition to sequence control of initiation through suitable TSS sequences, the probability of initiation across promoter positions is also controlled by multiple architectural features shown in **(A)**. These include the inhibition of initiation near a core promoter that diminishes as scanning proceeds ("distance restriction"), biochemical restrictions on how far scanning can proceed that increase as scanning proceeds ("scanning processivity"), and "Pol II flux", which represents the decrease in amount of scanning Pol II as scanning proceeds due to conversion of scanning Pol II to transcribing Pol II upon initiation. **(C)** Construction of promoter libraries that control TSS sequence context. Top panel shows schematic of the *SNR37* promoter and its TSS usage distribution based on TSS-seq (Chenxi Qiu et al., 2020). *SNR37* has a focused and highly efficient TSS region. Bottom panel shows schematic of the Pol II MASTER libraries used in this study. A duplication of the *SNR37* TSS region was inserted before native TSS region, and the -8 to +1 positions relative to native *SNR37* +1 TSS (black box) were replaced by a 9 nt randomized region to provide almost all possible sequences. The second *SNR37* TSS region functions as a "Flux Detector" (FD) to capture Pol II flux that fails to initiate within the randomized region and allow determination of initiation efficiency within the randomized region. A barcode region (purple box) allows RNA products to be assigned to respective promoter variants. The *GAL1* UAS allows control of library expression, and the *GFP* ORF and *CYC1* terminator support termination and stabilize RNA products. **(D)** TSS usage distributions at TSS and FD regions for different promoter variant "AYR", "BYR", and "ARY" libraries are shown on the left. TSS usages from designed +1 TSS and positions upstream are in red and grey, respectively. TSS usage from the FD region is in brown. The definition of "TSS efficiency" and overall TSS efficiency calculations for the aggregate +1 TSS of all in different libraries are shown in the middle. Example correlation plots of TSS efficiency calculations for +1 TSSs from individual promoter variants in Pol II MASTER libraries between representative biological replicates are show on right. Pearson r and number of compared variants are shown.

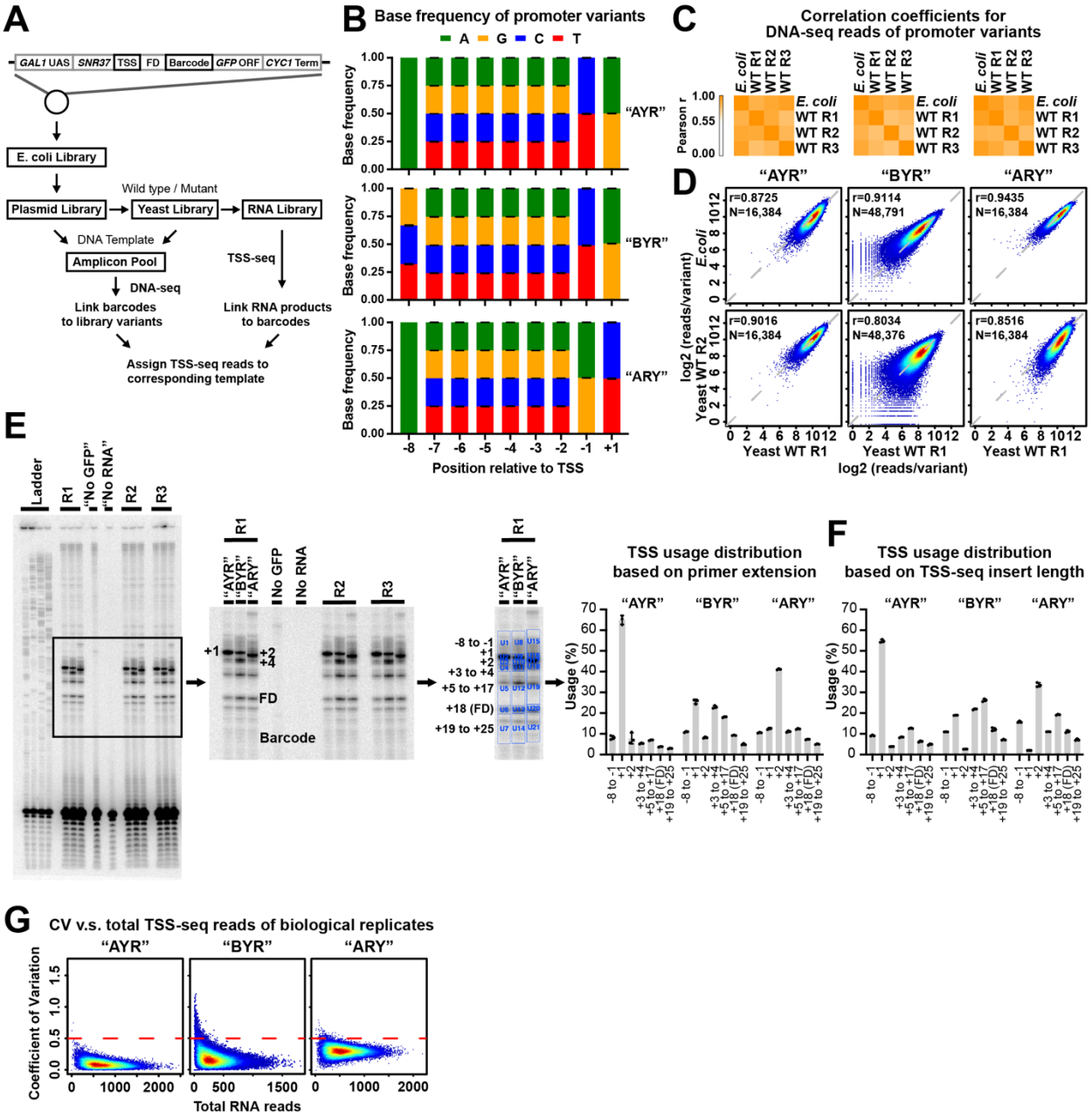


Figure S1. High level of reproducibility and coverage depth of library variants. **(A)** Schematic representation of experimental approach. Promoter libraries with almost all possible sequences within a 9 nt randomized region within a promoter context designed for specific functionalities were constructed on plasmids. Libraries were of three types, designated "AYR", "BYR", and "ARY" based on compositions of their randomized regions. Plasmids were amplified in *E.coli* and transformed into yeast strains with wild type or mutated Pol II. DNA and RNA were extracted from yeast pellets and prepared for DNA-seq and TSS-seq. **(B)** Base frequencies at positions within randomized region of promoter variants indicating unbiased coverage of randomized regions. Bars are mean +/- standard deviation of mean of promoter variants in WT and

four Pol II mutants. **(C)** Heatmap illustrating hierarchical clustering of Pearson correlation coefficients of reads per promoter variant for libraries amplified in *E.coli* and three biological replicates of these same libraries transformed into yeast. **(D)** Example correlation plots of DNA reads count of promoter variants for *E. coli* and yeast WT biological replicates. Pearson r and number of compared variants are shown. **(E)** Bulk primer extension for RNA produced from promoter variant libraries and quantification for biological replicates transformed into WT yeast. “No GFP” control used an RNA sample without library transformed. “No RNA” control used a sample of nuclease-free water. Dots represent three biological replicates. Bars are mean +/- standard deviation of mean. **(F)** TSS usage distribution based on insert length of TSS-seq reads generated from transformed libraries. Dots represent three biological replicates. Bars are mean +/- standard deviation of mean. Compare similarity to distribution in **(E)**. Note that primer extension will blur usage into adjacent upstream position due to some level of non-templated addition of C to RNA 5’ ends. **(G)** Heat scatter plots of Coefficient of Variation (CV, y axis) versus total RNA sequencing reads per promoter variant in each of three Pol II MASTER libraries. A cutoff of CV=0.5 was used to filter out variants with higher variance.

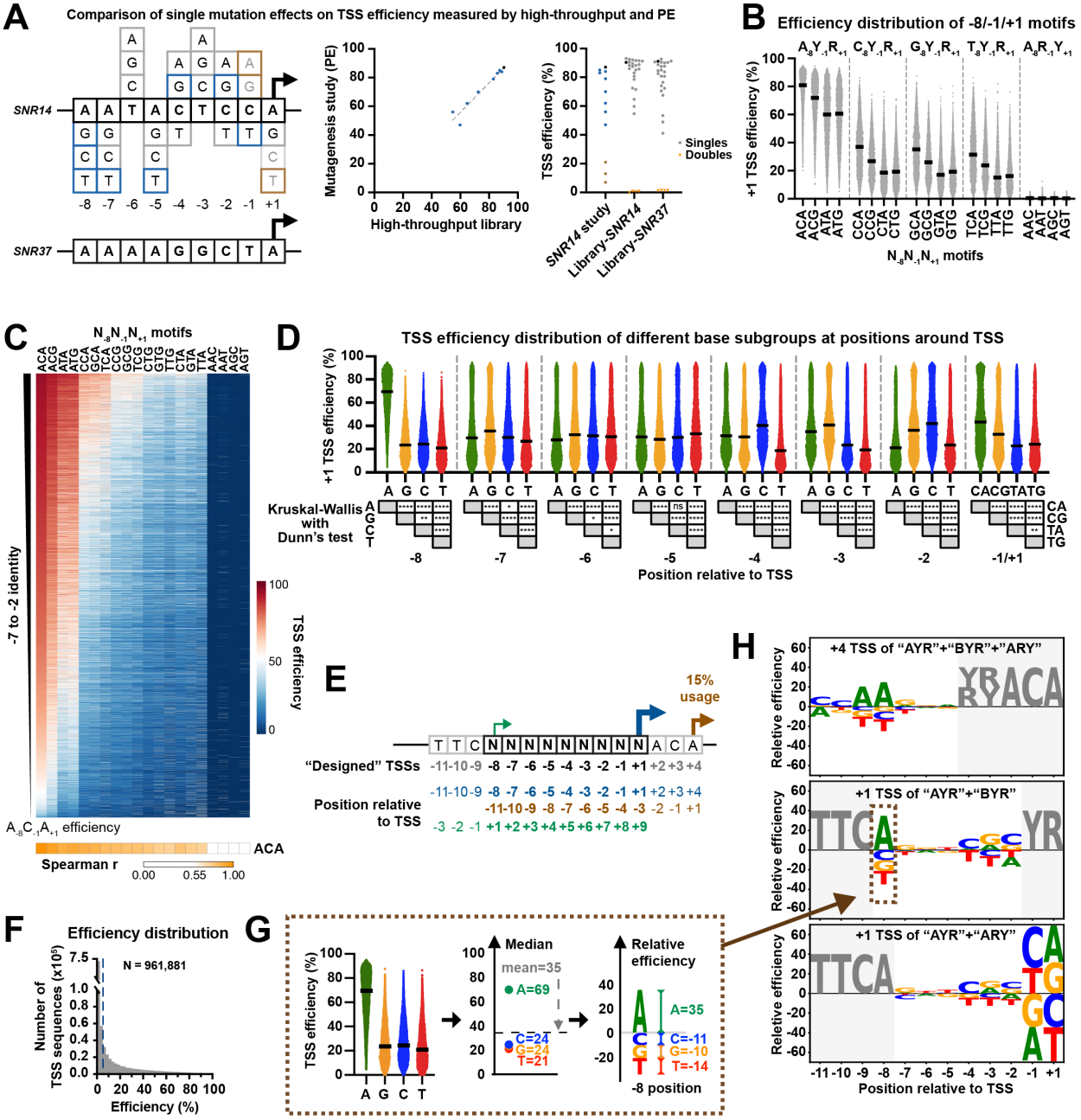


Figure 2. TSS sequence and its surrounding sequences modulate initiation efficiency over a wide range.

(A) Comparison of single mutation effects on TSS efficiency measured by high-throughput and primer extension. (Left) Sequences of *SNR14* and *SNR37* TSS regions (in black boxes, including positions between -8 to +1 relative to TSS) and all possible single substitutions of *SNR14* TSS region. Single substitutions included by both a prior *SNR14* mutagenesis study (Kuehner & Brow, 2006) and Pol II MASTER libraries are in blue while those lacking in our study are in brown. Additional substitutions analyzed here are in gray. (Middle) High correlation of TSS efficiency measured by Pol II MASTER and primer extension. Mutation classes color coded as on left. (Right) Range

of effects of single base substitutions on TSS efficiency for *SNR14*- and *SNR37*-related sequences. Mutation classes color coded as on left. Single substitutions absent from Pol II MASTER because of library design (R_{-1} and Y_{+1}) were expected to have super low TSS efficiencies. Double substitutions of *SNR14* and *SNR37* TSS region included in Pol II MASTER "ARY" library are shown as orange inverted triangles and show almost no efficiency. **(B)** Pol II initiation shows a strong preference for A_{-8} and $C_{-1}A_{+1}$ containing variants. All promoter variants were divided into 20 groups defined by bases at positions -8, -1 and +1 relative to the designed +1 TSS, and their +1 TSS efficiencies were plotted as spots. Lines represent median efficiencies of each group. **(C)** +1 TSS efficiency of all -7 to -2 sequences within each $N_{-8}N_{-1}N_{+1}$ motif in WT, rank ordered by efficiency of their $A_{-8}C_{-1}A_{+1}$ version, is shown as a heat map. The x-axis is ordered based on median efficiency values for each $N_{-8}N_{-1}N_{+1}$ motif group, as shown in **B**. Statistical analyses by Spearman's rank correlation test between $A_{-8}C_{-1}A_{+1}$ group and all groups are shown beneath the heat map. **(D)** Efficiency distributions of designed +1 TSSs grouped by base identities between -8 and +1 positions. Statistical analyses by Kruskal-Wallis with Dunn's multiple comparisons test for base preference at individual positions relative to +1 TSS are shown beneath data plots. Lines represent median values of subgroups. ****, $P \leq 0.0001$; ***, $P \leq 0.001$; **, $P \leq 0.01$; *, $P \leq 0.05$. **(E)** Definition of "designed" +1 TSS (designated as +1) and positions relative to this TSS (blue TSS arrow and sequence). TSS usage generated from upstream or downstream of "designed" +1 TSS (green and brown TSS arrows and sequences, respectively) allows us to study sequence preferences at positions -11 to -9 and +2 to +9 relative to site of initiation. **(F)** Histogram showing the distribution of measured efficiencies for all designed -8 to +4 TSSs of all promoter variants from "AYR", "BYR" and "ARY" libraries in WT. Dashed lines mark the 5% efficiency cutoff used to designate a TSS as active. Total number of TSS sequences is shown. **(G)** Schematic illustrating how "relative efficiency" is calculated and visualized as sequence logo in **H**. At a particular position relative to a particular TSS, first, all variants were divided into four base subgroups defined by base at this position. Next, median values of each base group were extracted and centered based on the mean of all median values. The centered median values were defined as "relative efficiencies", representing preferences for bases at this particular position. Finally, relative efficiencies of bases were visualized as sequence logos. Positive and negative values indicate relatively preferred or less preferred bases. **(H)** Pol II initiation shows distinct sequence preference at positions around the TSS. Top panel was generated using datasets of designed +4 TSS deriving from "AYR", "BYR" and "ARY" libraries. Middle panel was generated using datasets of designed +1 TSS deriving from "AYR" and "BYR" libraries. Bottom panel was generated using datasets of designed +1 TSS deriving from "AYR" and "ARY" libraries. Positions that contain fixed or not completely randomized bases are shown in grey.

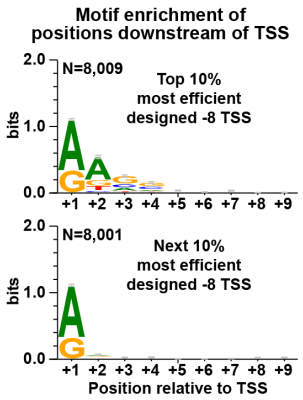


Figure S2. Sequence preference at TSS downstream positions
$A_{+2}G_{+3}G_{+4}$ motif enrichment is apparent for the top 10% most efficient designed -8 TSS. $A(/G)_{+2}G(/C)_{+3}G(/C)_{+4}$ motif enrichment was observed for the top 10% most efficient -8 TSSs but not for the next 10% most efficient TSSs. $A(/G)_{+1}$ enrichment observed for top 20% most efficient TSSs is consistent with the +1R preference of TSS. Numbers (N) of variants assessed are shown.

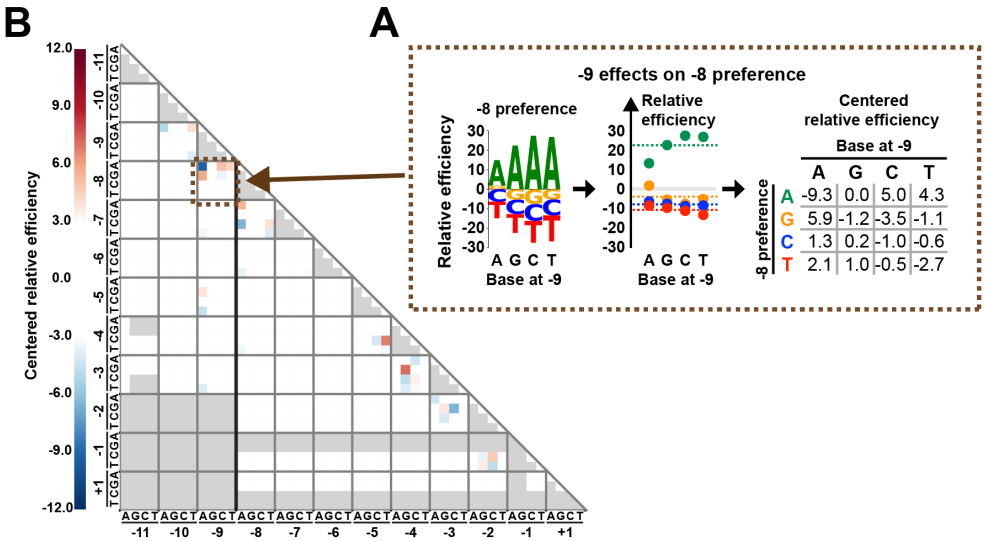


Figure 3. Sequence interaction between positions during TSS initiation
(A) Schematic illustrating how sequence interaction between positions is calculated and visualized as a heat map in **B**. Using -9/-8 positions as an example, relative efficiencies at position -8 when different bases were present at position -9 were calculated. Next, relative efficiencies of each base were centered based on the mean of all relative efficiencies of a particular base. Finally, the centered relative efficiencies matrix was visualized as a heat map to represent the interaction between examined positions. **(B)** Sequence interactions are mainly observed at neighboring positions. Red and blue indicate positive and negative interactions, respectively. Missing values are shown in grey. Interactions related to positions -11 to -9 were calculated using datasets of designed +4 TSS deriving from “AYR”, “BYR” and “ARY” libraries. Other interactions

were calculated using datasets of designed +1 TSS deriving from “AYR” and “BYR” libraries.

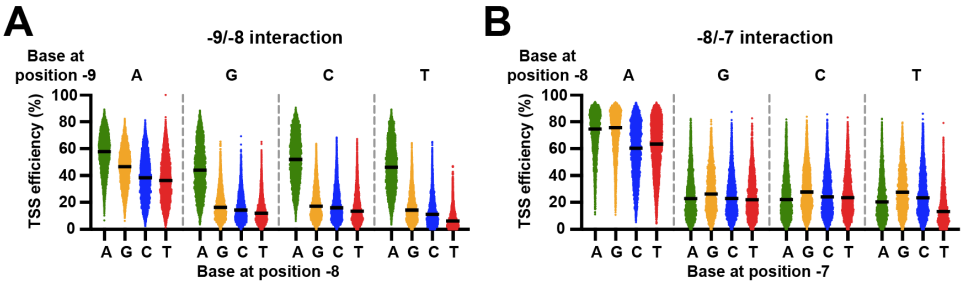


Figure S3.
(A) An A at position -9 results in different sequence preferences at position -8. The dataset of designed +4 TSSs deriving from “AYR”, “BYR” and “ARY” libraries was used to detect -9/-8 interaction. All variants were divided into 16 subgroups defined by bases at positions -9 and -8 relative to designed +4 TSS, and then their TSS efficiencies were plotted. Lines represent median values of subgroups. **(B)** An A at position -8 results in different sequence preferences at position -7. The dataset of designed +1 TSSs deriving from “AYR” and “BYR” libraries was used to detect -8/-7 interaction. Calculations same as -9/-8 interaction described in **A**.

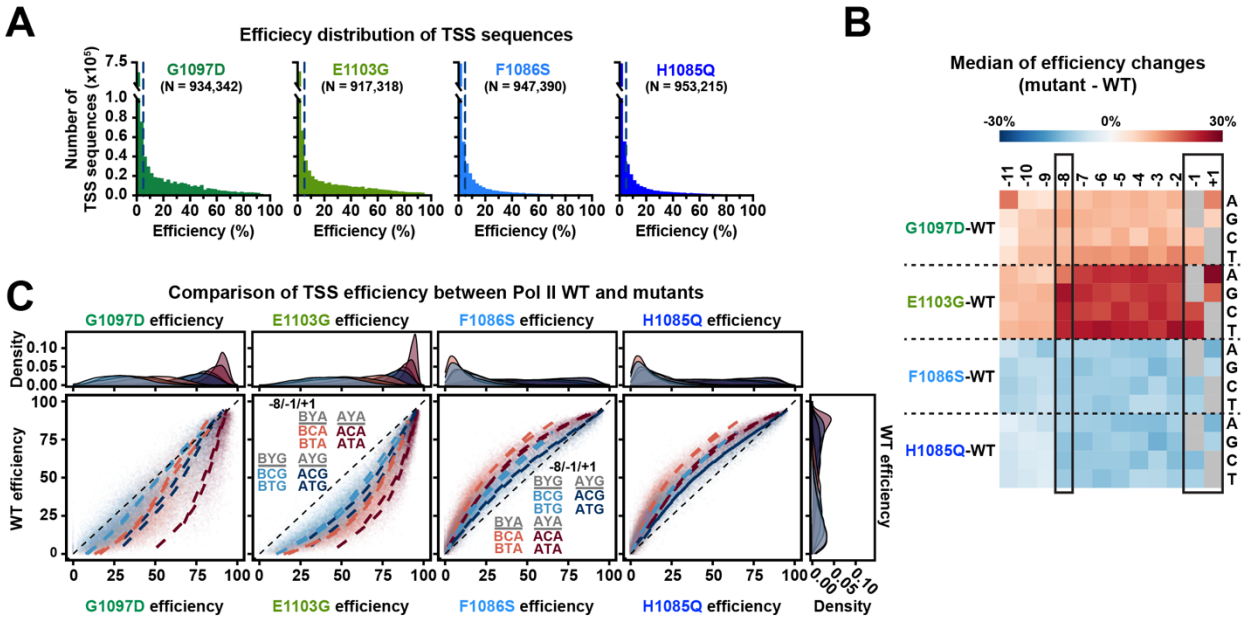
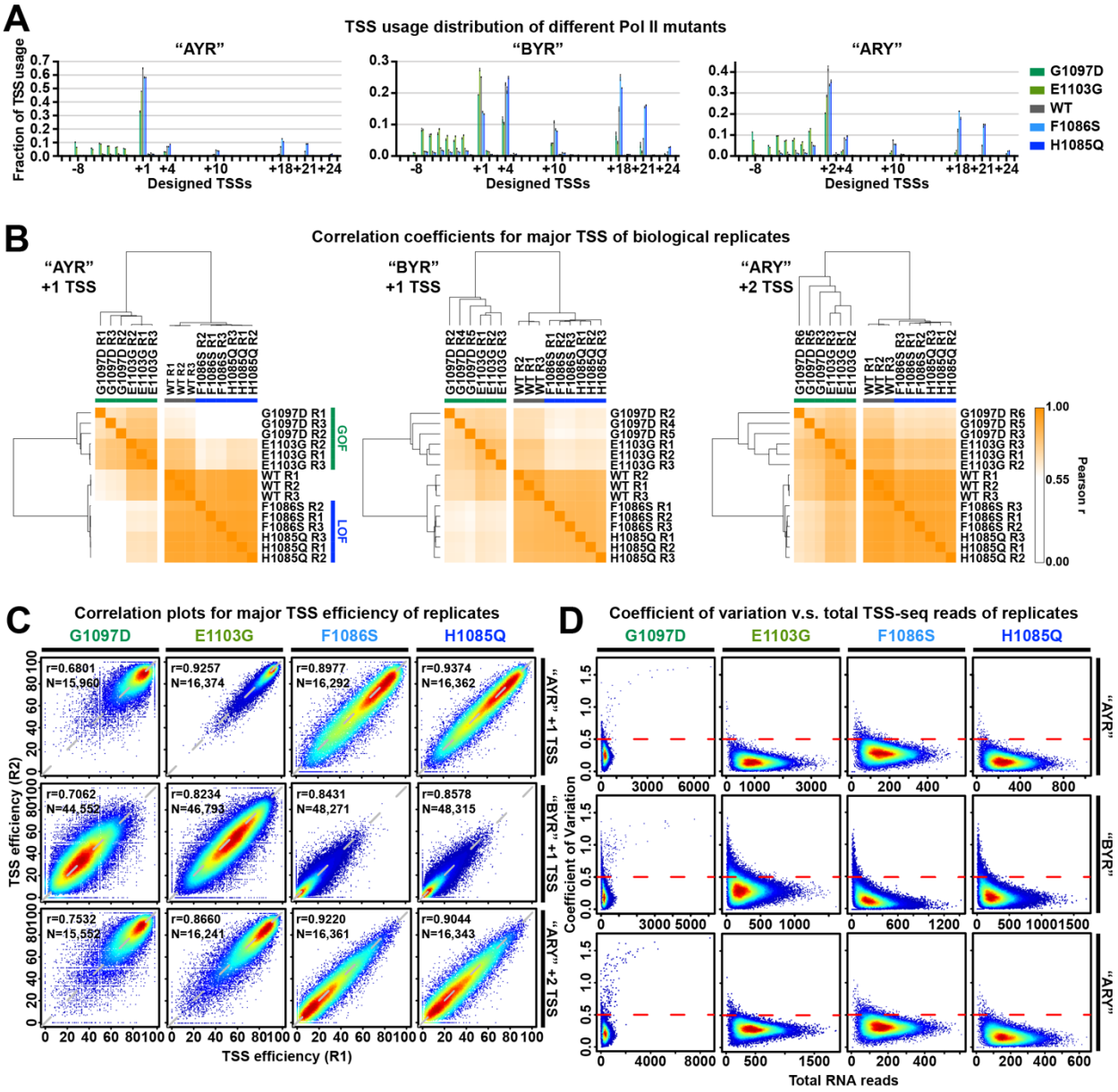
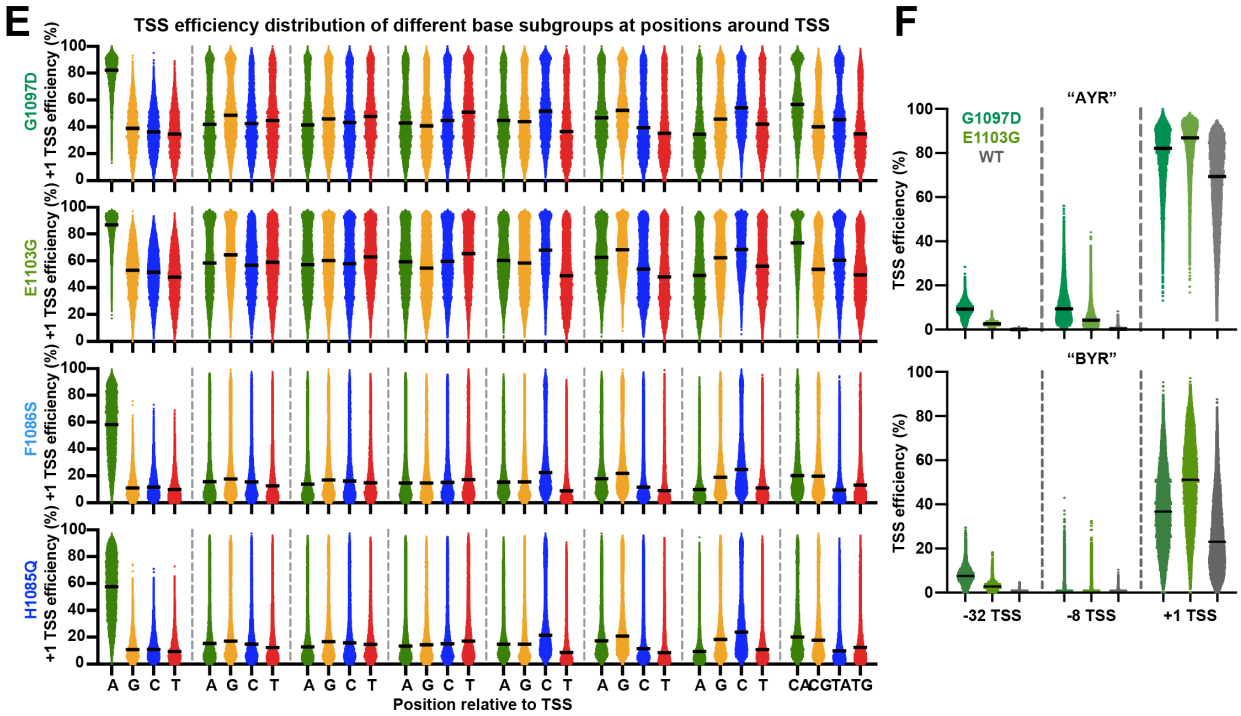
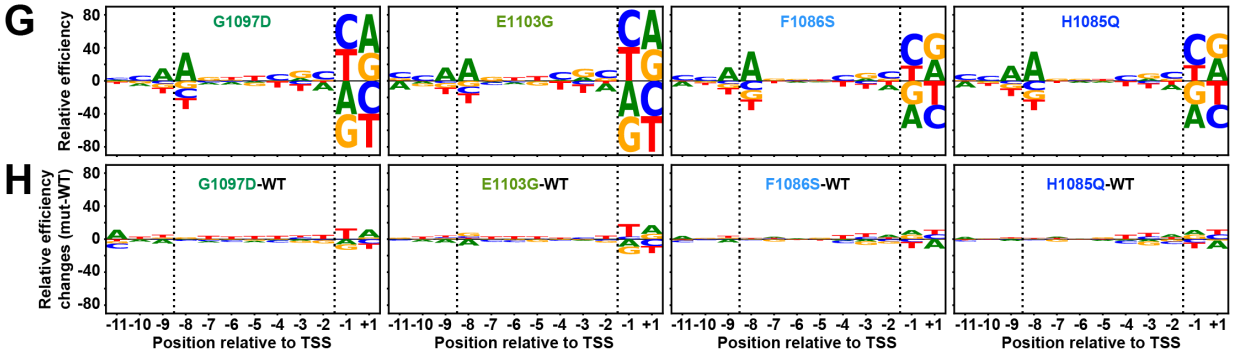


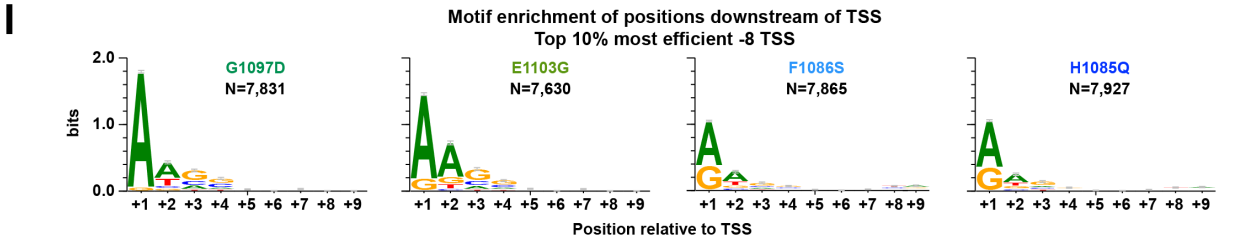
Figure 4. Pol II mutants alter TSS efficiency for all possible TSS motifs while showing selective effects for base at +1.
(A) Histograms showing the distribution of measured efficiencies for all designed -8 to +4 TSSs of all promoter variants deriving from “AYR”, “BYR” and “ARY” libraries in Pol II mutants. Dashed lines mark the 5% efficiency cutoff used to designate a TSS as active. Total numbers of TSS sequences are shown. **(B)** Pol II mutants alter TSS efficiencies

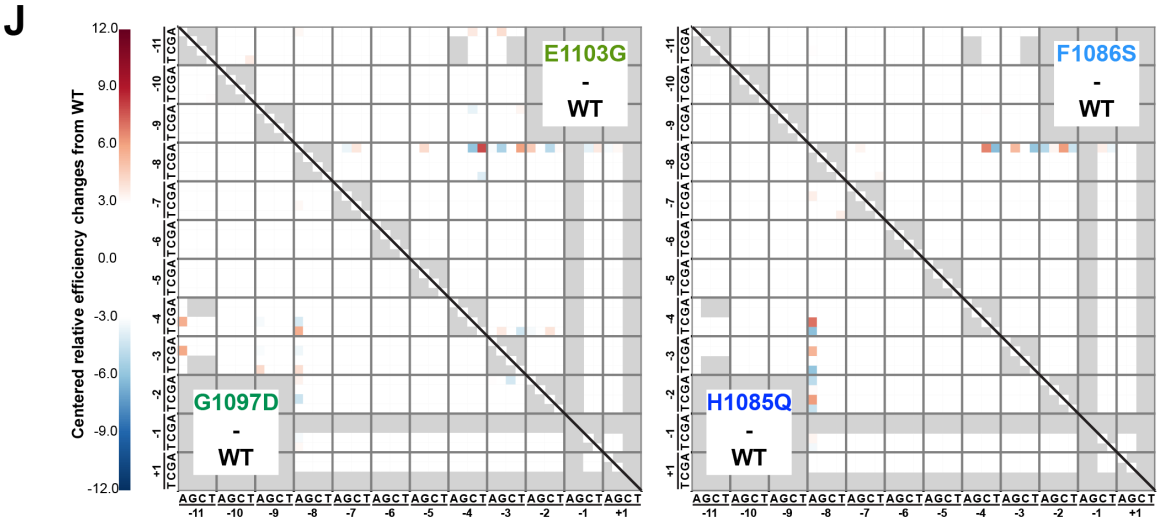
across all motifs, corresponding to direction of change to Pol II catalytic activity *in vitro*. TSS efficiency changes for each TSS variant were first determined by subtracting WT efficiency from Pol II mutant efficiency. The medians of efficiency changes for variant groups with indicated bases at each position relative to TSS were then calculated and illustrated in a heat map. Positive (red) values indicate Pol II mutants increased overall efficiency while negative (blue) values indicate decreased overall efficiency. **(C)** WT TSS efficiency for TSS variants divided into motif groups are plotted for mutant TSS efficiencies for the same TSS groups. The eight possible groups of TSSs for A/B_{-8}C/T_{-1}A/G_{+1} motifs were plotted and curve fit. Histograms show density of variants within each -8/-1/+1 subgroups. As to position -8, A_{-8} containing motifs show higher efficiency than B_{-8} containing motifs in both Pol II GOF (G1097D and E1103G) and LOF (F1086S and H1085Q) mutants (A_{-8} motifs: maroon and blue vs B_{-8} motifs: light coral and light blue). This is consistent with the proposed function of -8A to retain TSSs longer in the Pol II active site during scanning. This means that -8A may boost the positive effects by GOF mutants, therefore Pol II GOF mutants showed greater effects on A_{-8} motifs compared to B_{-8} motifs. In contrast, -8A compensates for active site defects of LOF mutants, therefore Pol II LOF mutants showed reduced effects on A_{-8} motifs compared to B_{-8} motifs. Both GOF and LOF mutants show reduced effects on G_{+1} motifs relative to A_{+1} motifs (G_{+1} motifs: light coral and maroon vs A_{+1} motifs: light blue and blue).


A
TSS usage distribution of different Pol II mutants
"AYR"
"BYR"
"ARY"
Fraction of TSS usage
Designed TSSs
G1097D
E1103G
WT
F1086S
H1085Q
B
Correlation coefficients for major TSS of biological replicates
"AYR" +1 TSS
"BYR" +1 TSS
"ARY" +2 TSS
GOF
LOF
Pearson r
C
Correlation plots for major TSS efficiency of replicates
G1097D
E1103G
F1086S
H1085Q
r=0.6801 N=15,960
r=0.9257 N=16,374
r=0.8977 N=16,292
r=0.9374 N=16,362
r=0.7062 N=44,552
r=0.8234 N=46,793
r=0.8431 N=48,271
r=0.8578 N=48,315
r=0.7532 N=15,552
r=0.8660 N=16,241
r=0.9220 N=16,361
r=0.9044 N=16,343
"AYR" +1 TSS
"BYR" +1 TSS
"ARY" +2 TSS
TSS efficiency (R2)
TSS efficiency (R1)
D
Coefficient of variation v.s. total TSS-seq reads of replicates
G1097D
E1103G
F1086S
H1085Q
"AYR"
"BYR"
"ARY"
Coefficient of Variation
Total RNA reads


E
TSS efficiency distribution of different base subgroups at positions around TSS
G1097D
E1103G
F1086S
H1085Q
+1 TSS efficiency (%)
A G C T A G C T A G C T A G C T A G C T A G C T A G C T CACGTATG
Position relative to TSS
F
"AYR"
"BYR"
G1097D
E1103G
WT
TSS efficiency (%)
-32 TSS
-8 TSS
+1 TSS


G
H
G1097D
E1103G
F1086S
H1085Q
Relative efficiency
G1097D-WT
E1103G-WT
F1086S-WT
H1085Q-WT
Relative efficiency changes (mut-WT)
Position relative to TSS


I
Motif enrichment of positions downstream of TSS
Top 10% most efficient -8 TSS
G1097D
N=7,831
E1103G
N=7,630
F1086S
N=7,865
H1085Q
N=7,927
bits
Position relative to TSS



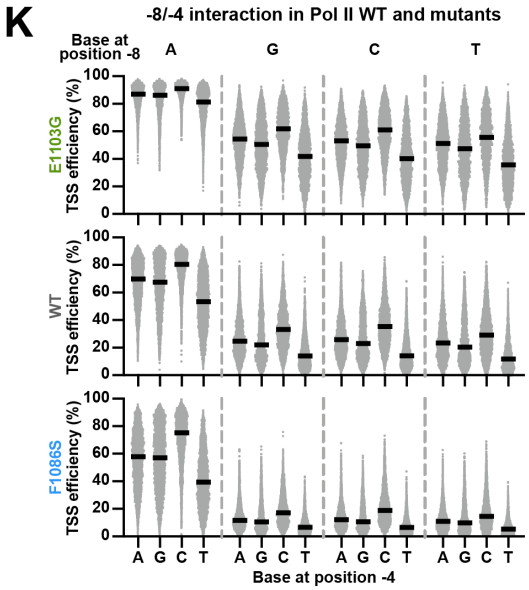


Figure S4. Pol II activity mutant effects on TSS selection

(A) TSS usage distributions at designed -10 to +25 TSSs in different promoter variant "AYR", "BYR", and "ARY" libraries. Pol II GOF and LOF mutants shifted TSS usage upstream or downstream relative to WT, respectively. Dots represent three biological replicates. Bars are mean +/- standard deviation of the mean. **(B)** Hierarchical clustering of Pearson correlation coefficients of TSS efficiencies for major TSSs (designed +1 TSS for "AYR" and "BYR" libraries, +2 TSS for "ARY" library) for three biological replicates for WT or mutant Pol II illustrated as a heat map. **(C)** Example correlation plots of TSS efficiency of major TSSs of promoter variants between representative biological replicates. Pearson r and number of compared variants are shown. **(D)** Plots of the coefficient of variation (CV) versus the total RNA reads for three yeast replicates in Pol II mutants. The red dashed lines mark the CV=0.5 cutoff, which was chosen as an arbitrary cutoff for variants showing reasonable reproducibility across three biological

replicates. G1097D replicates contain outliers because these slow growing strains are susceptible to genetic suppressors. These outliers are filtered out using our CV cutoff because of high CVs caused by suppressor only existing in one of three biological replicates. **(E)** TSS efficiency distributions of designed +1 TSSs of Pol II mutants for base subgroups at individual positions relative to +1. Identical analysis as in **Figure 2D** for WT was performed for Pol II mutant libraries. **(F)** Pol II GOF G1097D showed greater increase in efficiency than GOF allele E1103G at upstream TSSs (designed -32 and -8 TSSs), while E1103G showed stronger effects at designed +1 TSS than G1097D. This may indicate that where Pol II is in the scanning process may affect efficiency. **(G)** Pol II initiation sequence preference in Pol II mutants. Identical analysis as in **Figure 2H** for WT was performed for Pol II mutant libraries. Sequence logos were generated using multiple datasets and the dashed lines indicate divisions in datasets used to generate them. Specifically, preferences at positions -11 to -9 were generated using datasets of designed +4 TSS deriving from “AYR”, “BYR” and “ARY” libraries. Preferences at positions -8 to -2 were generated using datasets of designed +1 TSS deriving from “AYR” and “BYR” libraries. Preferences at positions -1 and +1 were generated using datasets of designed +1 TSS deriving from “AYR” and “ARY” libraries. **(H)** Pol II catalytic mutants do not change overall sequence specificity. Relative efficiency changes of mutants were calculated by subtracting relative efficiency in WT from that in mutants, followed by being visualized as sequence logo. The apparently changed preference at positions -1 and +1 could be biased by a ceiling effect for $Y_{-1}R_{+1}$ and a floor effect for $R_{-1}Y_{+1}$. This means that TSS efficiencies of $Y_{-1}R_{+1}$ containing variants have much less room to be increased by GOF mutants compare to $R_{-1}Y_{+1}$ containing variants. In contrast, TSS efficiencies of $Y_{-1}R_{+1}$ containing variants have much more room to be decreased by LOF mutants compared to $R_{-1}Y_{+1}$ containing variants. **(I)** Motif enrichment for top the 10% most efficient -8 TSSs for Pol II mutants. Identical motif enrichment analysis as in **Figure S2** top panel for WT was performed to Pol II mutant libraries. Numbers (N) of variants assessed are indicated. **(J)** Pol II mutants do not strongly affect interactions between promoter positions. Heatmaps show Pol II mutant effects on sequence interactions between promoter positions, as determined by subtracting centered relative efficiency in WT from that in mutants. Apparently altered -8A related interactions are biased by different levels of innate TSS efficiency and ceiling or floor effects, which are addressed in **K**. **(K)** Pol II mutants do not likely alter the -8/-4 interaction. Identical interaction analysis as in **Figure S3A** for the -9/-8 interaction in WT was performed to positions -8 and -4 using dataset of designed +1 TSS deriving from “AYR” and “BYR” libraries in WT, Pol II E1103G or F1086S libraries. No obvious interactions were observed between positions -8 and -4 in WT (**Figure 3B**). The consistent relative preferences at position -4 (C>A>=G>T) in WT and Pol II mutants suggested Pol II mutants did not cause novel -8/-4 interaction. Instead, the apparent interaction (changed centered relative efficiency observed in (**J**)), was due to range effects, namely compression of efficiency at the low and high ends of the scale.

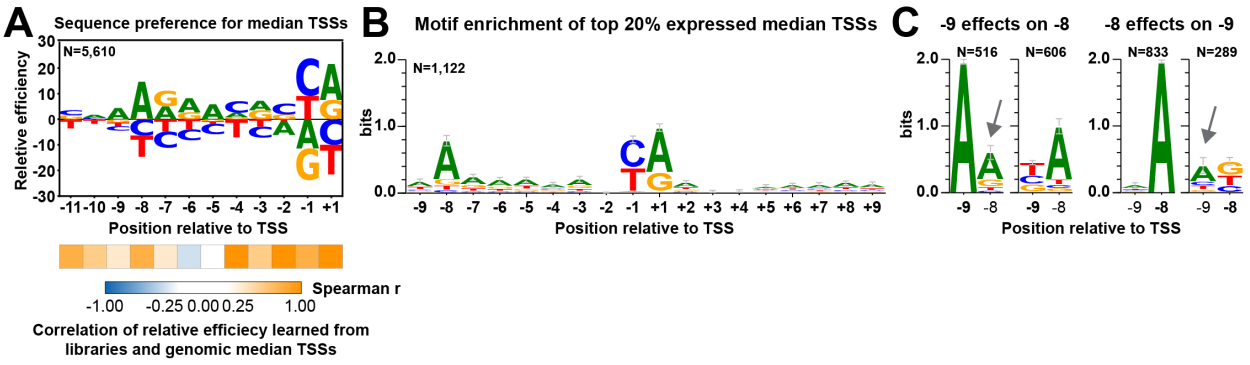


Figure 5. Learned initiation preferences are predictive of TSS efficiencies at genomic promoters

(A) Sequence preference determined from genomic median TSSs are congruent with library determined TSS preferences, except there is genomic enrichment for A at positions -7 to -5. Sequence context and TSS efficiency of genomic median TSSs were extracted from genomic TSS-seq data (GSE182792) (Zhao et al., 2021). Calculation and visualization were as performed for promoter variant libraries. The number (N) of genomic median TSSs examined is shown. Statistical analyses by Spearman's rank correlation test between relative efficiency at individual positions learned from promoter variant libraries and genomic median TSSs are shown beneath the sequence logo. **(B)** A-richness upstream and downstream of TSS is observed for highly expressed median TSSs. The number (N) of analyzed median TSSs is shown. **(C)** Having an A at either position -9 or -8 reduces enrichment of A at the other position. Top 20% expressed median TSSs were divided into subgroups based on bases at position -9 or -8. Motif enrichment analysis was individually performed to subgroups. Numbers (N) of median TSSs within each subgroup are shown.

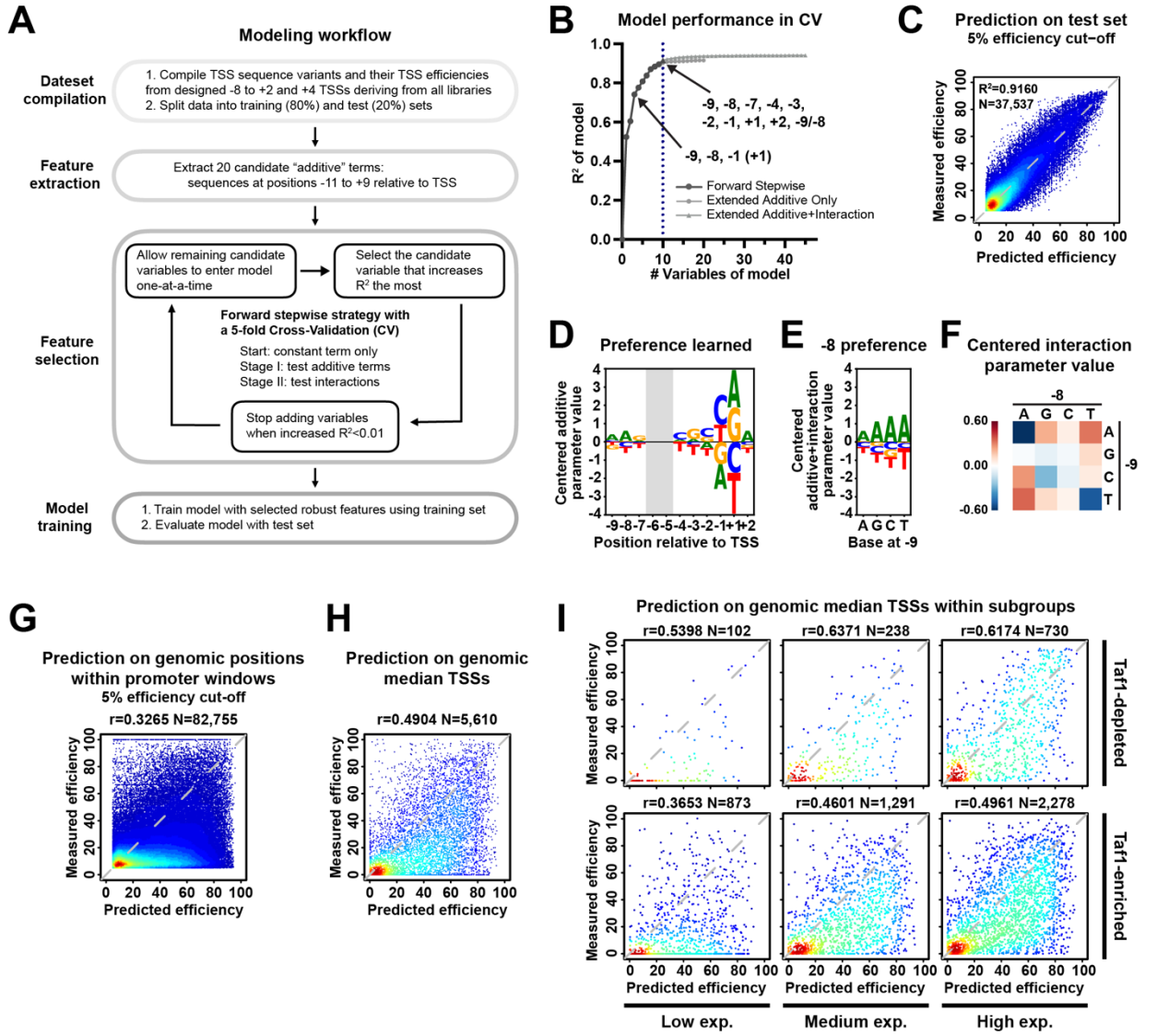


Figure 6. Logistic regression models of TSS sequence function.
(A) Overview of modeling process. (1) Variants including designed -8 to +2 and +4 TSSs deriving from "AYR", "BYR" and "ARY" libraries with available TSS efficiency were pooled for modeling. (2) Sequences at positions -11 to +9 relative to TSS of each variant were extracted. (3) To identify robust features, a forward stepwise selection strategy coupled with a 5-fold cross-validation for logistic regression was used. Data were randomly split into training (80%) and test (20%) sets. The training set was used for a stepwise regression approach that starts from a model with a constant term only and adds variables that improve the model the most one at a time, until a stopping criterion is met. In stage I, additive terms (sequences at positions -11 to +9) were tested. In stage II, interactions between positions selected in stage I were tested. Model performance was evaluated with R^2. The stopping criterion for adding additional variables was an increase $R^2 < 0.01$. (4) A logistic regression model containing selected robust features was trained with training set and then evaluated with the test set. **(B)** Regression modeling identifies key DNA sequences and interactions contributing to

TSS efficiency. Nine additive parameters plus one interaction were selected for the final model. Dots represents average R^2 obtained in a 5-fold Cross-Validation (CV) strategy for logistic regression models using different numbers of features. The black line with SD error bars represents models with the best performance under a certain number of predictors. Note that model with as few as three additive parameters could explain 74.10% of TSS efficiency variation in our WT libraries. **(C** to **F)** Good performance of model including sequences at nine positions and the -9/-8 interaction indicates that TSS efficiency in our libraries is mainly regulated by the included features. (**C**) A scatterplot of comparison of measured and predicted efficiency of test sets, with a 5% efficiency cut-off. Model performance R^2 on entire test and number (N) of data points shown in plot are shown. (**D**) A sequence logo of centered additive parameters. The coefficients for bases at a particular position was centered and visualized as a sequence logo. (**E**) A sequence logo showing learned preference at position -8 when different bases existing at position -9, with -9/-8 interaction included. The -9/-8 interaction parameters were added to corresponding additive coefficients for bases at position -8. The additive plus interaction parameters were then centered and visualized as sequence logos. (**F**) A heat map of centered parameters for -9/-8 interaction illustrating how bases at one position affect preference at another position. **(G** to **H)** Efficiency prediction for positions within known promoter windows in WT shows overall over-prediction. Scatterplots of comparison of measured and predicted TSS efficiencies of all positions (with a 5% efficiency cut-off) (**G**) or median TSSs (**H**) within 5979 known genomic promoter windows (Chenxi Qiu et al., 2020) with available measured efficiency. **(I)** Model shows better performance on Taf1-depleted promoters and promoters with medium to high expression. Scatterplots of comparison of measured and predicted TSS efficiency of median TSSs subgrouped by promoter classes and expression levels. Expression levels of genomic promoters are defined based on their total TSS-seq reads in the promoter window in the examined datasets: low, [0, 200); medium, [200, 1000); high, [1000, max). Pearson r and number (N) of compared variants are shown.

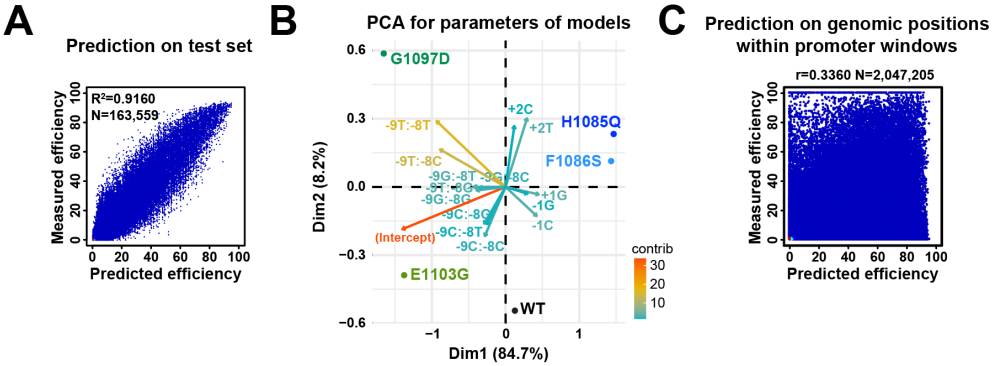


Figure S5.
(A) A scatterplot of comparison of measured efficiencies and predicted efficiencies within the test sets. Model performance R^2 on entire test set and number of data points shown in plot are shown. **(B)** PCA analysis for parameters of models trained by WT and Pol II mutant datasets. The top 15 contributing variables are shown. GOF and LOF mutants were separated from WT by the 1st principal component. GOF G1097D and

E1103G were further distinguished by 2nd principal component by additional position +2 information, which is consistent with results in **Figure S4I**, where G1097D and E1103G differentially altered +2 sequence enrichment. **(C)** A scatterplot of comparison of measured and predicted TSS efficiencies of all positions within 5979 known genomic promoter windows (Chenxi Qiu et al., 2020) with available measured efficiency. Pearson r and number (N) of compared variants are shown. Most promoter positions (82%, 1,678,406 out of 2,047,205) showed no observed efficiency, which is expected because TSSs need to be specified by a core promoter and scanning occurs over some distance downstream.

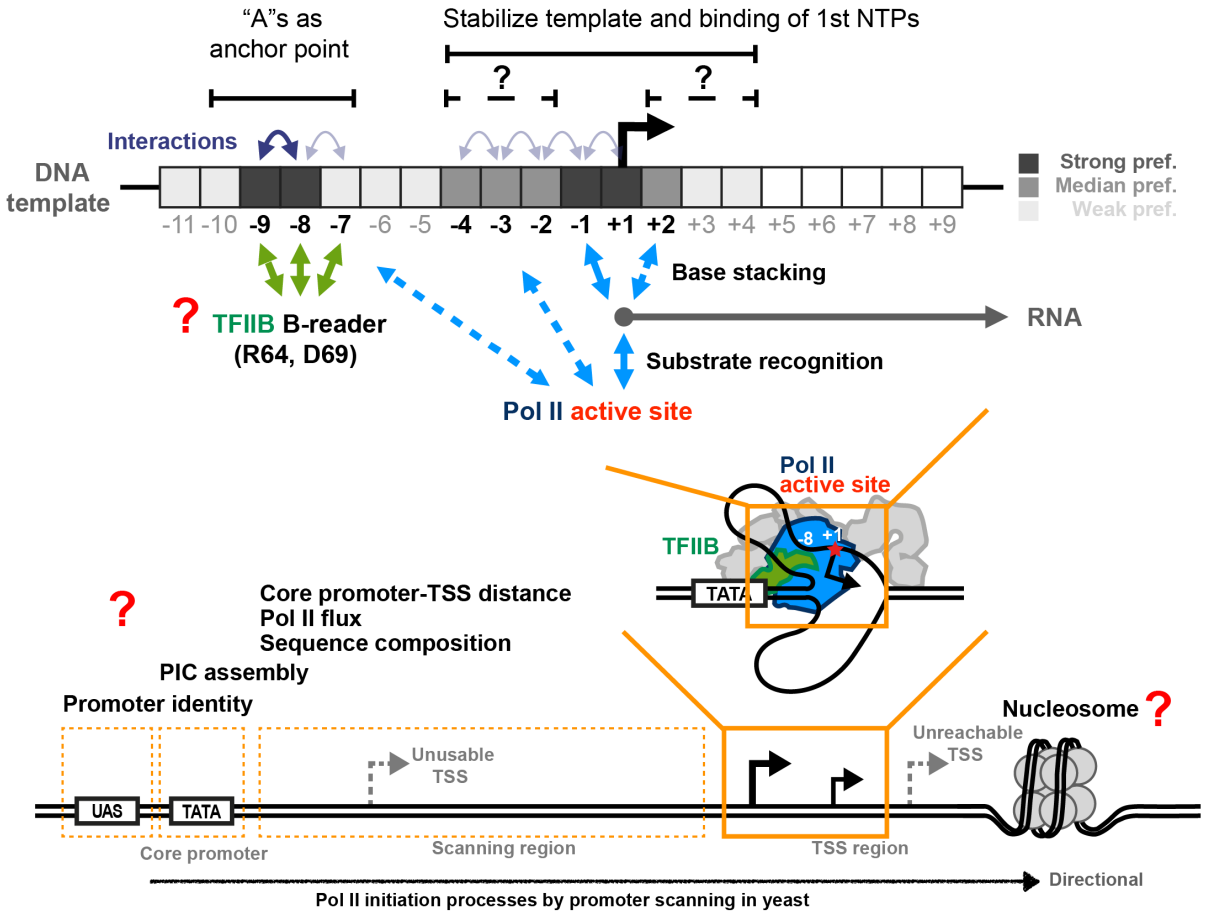


Figure 7. Model for TSS sequence preference regulated by multiple mechanisms. Top panels shows determined contribution of sequence at positions around TSS and proposed corresponding mechanism by this study. Two major groups of positions around TSS contribute to TSS selection: bases around TSS (actual initiating site) and bases around position -8. The TSS and adjacent bases interact with Pol II active site, the 1st NTP or each other to facilitate stable binding of 1st NTP thus stimulate RNA synthesis. -8 and -9 Ts on the template strand with an additional interaction between -8 and -7 template strand positions are proposed to serve as an anchor point interacting with TFIIB B-reader domain allowing pausing of scanning and promotion of Pol II initiation at TSSs a fixed distance downstream. These preferences are reflection as As if

the analysis is on the coding strand. Positions and interactions that were identified by regression modeling as robust features are labelled in bold. Bottom panel shows other architectural features involved in Pol II transcription initiation likely additionally contributing to TSS selection and initiation efficiency that will be accessible to Pol II MASTER analysis.