Quantitative analysis of transcription start site selection in *Saccharomyces cerevisiae* reveals control by DNA sequence, RNA Polymerase II activity, and NTP levels

Yunye Zhu[1], Irina O. Vvedenskaya[2], Sing-Hoi Sze[3,4], 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
[3] Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX 77843, USA
[4] Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, USA

Abstract

Transcription start site (TSS) selection is a key step in gene expression and occurs at many promoter positions over a wide range of efficiencies. Here, we develop a massively parallel reporter assay to quantitatively dissect contributions of promoter sequence, NTP substrate levels, and RNA polymerase II (Pol II) activity to TSS selection by “promoter scanning” in *Saccharomyces cerevisiae* (Pol II MAssively Systematic Transcript End Readout, "Pol II MASTER"). Using Pol II MASTER, we measure the efficiency of Pol II initiation at 1,000,000 individual TSS sequences in a defined promoter context. Pol II MASTER confirms proposed critical qualities of *S. cerevisiae* TSS -8, -1, and +1 positions quantitatively in a controlled promoter context. Pol II MASTER extends quantitative analysis to surrounding sequences and finds they tune initiation over a wide range of efficiencies. These results enable the development of a predictive model for initiation efficiency based on sequence. We show that genetic perturbation of Pol II catalytic activity alters initiation efficiency mostly independently of TSS sequence, but selectively modulates preference for initiating nucleotide. Intriguingly, we find that Pol II initiation efficiency is directly sensitive to GTP levels at the first five transcript positions and to CTP and UTP levels at the second position genome-wide. These results suggest individual transcripts will be differentially sensitive to alterations in NTP levels, representing a cryptic layer of potential regulation at the level of Pol II biochemical properties. The results establish Pol II MASTER as a method for quantitative dissection of transcription initiation in eukaryotes.

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 [1-11]. 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 [3,12,13], and therefore is widespread in different cell types [6], developmental processes [4,12,14,15], growth conditions [16], responses to environmental changes, and cancers [17-19].

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 [20-27]. The efficiency of initiation at a given position within a region allowing initiation depends on multiple features. Location relative to the core promoter region, with bases scanned from upstream to downstream from the core promoter, is critically important, as is 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), and this preference is near universal for RNAPs[1,10,11,16,21,22,28-44].

It has long been recognized that DNA sequences at positions beyond the -1/+1 also contribute to the efficiency of initiation by Pol II. However, the sequence preferences at these positions are difficult to determine from genomic usage alone given differences in context, 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 transcribed strand and a T on the template strand [1,16,21,34,35]. Furthermore, promoters with high expression and/or with initiation focused primarily in a single, efficient TSS tend to show additional A enrichment at positions -7 to -3 on the transcribed strand (T on the template strand) [1,16,21,45,46]. 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. Promoter architectural factors are especially apparent in yeast – where initiation occurs by promoter scanning – because TSSs are proposed to be selected in a polar fashion from upstream to downstream.

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

[22]. 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 with upstream TSSs having priority, regardless of innate TSS strength. Kuehner and Brow introduced the 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**). TSS efficiency is defined as the usage of a particular TSS normalized to the usage of that position and all downstream positions. Downstream TSS usage is used to estimate the flux of scanning Pol II that scans past a particular TSS. This measure is essential for comparison of TSSs embedded in different promoter contexts and different positions relative to other TSSs.

Imbalanced promoter sequence distributions imposed by evolutionary processes also limit the 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 the enrichment for T on the transcribed strand upstream of the median TSS position (reflecting the middled of the TSS distribution) and A on the transcribed strand downstream of the median TSS position at highly expressed promoters, and a paucity of G/C content in general [21,45-48]. Therefore, the biased distribution of bases in yeast promoters leads to a biased distribution of TSS sequence motifs, where preferred TSS motifs, due to increased probability of a -8A in additional to evolutionary selection for it, show enrichment downstream of the median TSS position, and less-preferred motifs show enrichment upstream of the median TSS position [21]. We have found that hyperactive Pol II mutants ("gain-of-function" or GOF) and hypoactive Pol II mutants ("loss-of-function" or LOF) were observed to shift TSS usage distributions upstream or downstream, respectively, with differences in aggregate usage of particular TSS motifs. Differences in usage due to biased promoter sequence distribution cannot be disentangled from actual differences in TSS efficiencies. Moreover, other properties such as the biochemistry of transcription itself due to availability of NTP substrates, the biochemical properties of scanning processivity [20,25,27], promoter identity [49-51], or promoter chromatin could also contribute to initiation output [52]. Together, all of these factors would make contributions 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) [42,53-56], 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 and upon manipulation of NTP levels. We show that this system enables determination of the interface between initiation factor activity, transcription substrates, 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

126 positions. We find that Pol II mutant classes increase or decrease initiation efficiencies
127 for all possible sequences. This result is consistent with predictions that the primary
128 effects of altered TSS selection (directional shifts in TSS distributions) are driven by
129 initiation efficiency changes across all sequences and not on TSS sequence selection
130 per se. Our results further show that Pol II activity level does contribute to selective
131 efficiency of initiation at sequences +1A vs +1G and find that initiation efficiency is
132 selectively sensitive to individual NTP levels across genomic TSSs. Our findings
133 demonstrate that Pol II MASTER provides a quantitative platform for how initiation factor
134 activity together with promoter sequence contribute to Pol II transcription initiation in
135 vivo.

136 Results

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

139 TSSs are identified at yeast promoters by scanning from upstream near the core-
140 promoter to downstream (**Figure 1A**). While promoters are melted starting around +20
141 from the TATA box (if present), there is a distance restriction that reduces usage of
142 TSSs until they are within a region ~40-140 nt downstream from the core promoter [57]
143 (**Figure 1B**). Processivity of the scanning process, likely determined by TFIIH activity, is
144 expected to limit TSS usage on the downstream edge of promoters (**Figure 1A**,
145 "unreachable TSS"). Once initiation happens at any site within this window of
146 opportunity for initiation (**Figure 1A, B**), Pol II "flux" – the amount of Pol II progressing to
147 TSSs downstream – is necessarily reduced. Given these potential variables in TSS
148 context and the "first come-first served" basis of promoter scanning, determining how
149 some TSSs are stronger than others is difficult if only considering TSS usage at
150 genomic promoters. In order to specifically dissect how TSS sequence controls initiation
151 efficiency and how sequence interacts with Pol II catalytic activity or NTP levels in a
152 controlled context, we have established a massively parallel promoter variant assay "Pol
153 II MASTER" where we have embedded almost all possible sequences within a 9 bp
154 randomized TSS region (**Figure 1C**) constructed on plasmids and introduced into yeast
155 strains with wild type (WT) or mutated Pol II. The sequence libraries constructed are
156 illustrated in **Figure 1C** and are referred to by their base compositions relative to the
157 transcribed strand unless otherwise specifically noted. The libraries are named by
158 identities of bases at positions -8, -1, and +1. The "AYR" library has composition A_{-}
159 $_{8}NNNNNNY_{-1}R_{+1}$ (N=A, C, G, or T, Y=C or T, R=A or G) relative to the transcribed
160 strand, with "BYR" having composition $B_{-8}NNNNNNY_{-1}R_{+1}$ (B=C, G, or T), *etc*. Our three
161 libraries comprise 81,920 promoter variants, allowing up to 983,040 individual TSS
162 sequences to be evaluated, as each promoter variant may allow initiation from different
163 positions within or right after the randomized region.

164 This randomized TSS region was inserted into a controlled promoter context containing
165 specific functionalities (**Figure 1C**). <u>First</u>, the *GAL1* UAS was utilized to allow
166 expression control of libraries. <u>Second</u>, the TATA-box-to-TSS region of *SNR37* gene
167 was used as a "scanning region" to direct initiation within the downstream randomized

168 TSS region. This is because almost no RNA 5′ ends are observable from this scanning
169 region within its normal promoter context. Third, the native, highly-efficient *SNR37* TSS
170 region was inserted downstream of the randomized TSS region as a "Flux Detector"
171 (FD). Here, we employ the approach of Kuehner and Brow that a highly efficient
172 initiation region placed downstream an upstream TSS region may capture any
173 polymerases that happen to scan past the randomized TSS region. As a metric to
174 compare strength of different TSSs, we employ "TSS efficiency" as defined by Kuehner
175 and Brow [22], which measures the usage of a TSS relative to usage at that TSS and all
176 downstream positions (**Figure 1C**). This metric serves two key purposes: it allows
177 upstream and downstream starts to be compared as it takes priority effects into account
178 (upstream TSSs reduce the amount of polymerases scanning to downstream sites), and
179 it allows normalization of promoter usage using RNA only, engendering comparison
180 across promoters and libraries. Fourth, a 24 bp DNA barcode containing 20 positions of
181 randomized bases and 4 interspersed fixed bases (to exclude low-complexity
182 sequences) allows RNA products to be assigned to respective promoter variants. An
183 RNA barcode is critical as bases upstream of the TSS will not be present in the
184 transcribed RNA yet still contribute to TSS efficiency. Fifth, the *GFP* coding region and
185 *CYC1* terminator were added to support termination and stabilize RNA products.
186 Libraries were constructed by PCR sewing followed by cloning into a plasmid backbone.
187 After amplification in *E. coli*, plasmid libraries were transformed into Pol II WT and
188 mutant yeast strains in triplicate. Library expression was induced by addition of
189 galactose to the medium (4% final) for three hours. Both plasmid DNA and RNA
190 products were extracted from harvested yeast cells and amplified for DNA-seq and
191 TSS-seq (**Figure S1A**).

192 Several measures indicate high level of reproducibility and coverage depth of library
193 variants (**Figure 1D, Figure S1**). Base coverage in the randomized region was highly
194 even (**Figure S1B**). Correlation analysis for DNA-seq reads per promoter variant
195 suggested that yeast transformation did not alter promoter variant distribution (**Figure**
196 **S1C, D**). Bulk primer extension of libraries illustrated their average behavior and the
197 amount of initiation deriving from the randomized region (**Figure S1E**). As designed,
198 only a very small fraction of initiation was generated from the barcode region or further
199 downstream, validating that the flux detector captured scanning polymerases (**Figure**
200 **S1E, F**). Aggregate distribution of reads in our three libraries shows that as TSSs
201 decreased in efficiency from the most efficient library ("AYR") to least ("ARY") reads
202 shift from the designed +1 TSS to downstream positions (**Figure 1D**, left). The apparent
203 shift of TSS usage to position -1 in the ARY library was because the purine at the
204 designed -1 position serves as a +1 for newly created TSSs. **Figure 1D** (middle)
205 illustrates the aggregate TSS efficiency of each library based on usage at the designed
206 +1 TSS relative to usage at that position plus all downstream usage. Correlation
207 analysis for efficiency of library "major" TSSs (designed +1 TSS of "AYR" and "BYR"
208 libraries, +2 TSS of "ARY" library) demonstrated that biological replicates were highly
209 reproducible (**Figure 1D**, right). Therefore, we summed reads from the three biological
210 replicates, keeping TSSs that contained at least five TSS-seq reads in each replicate
211 and whose Coefficient of Variation (CV) in TSS-seq reads across replicates was less

212 than 0.5 (as a proxy for reproducible behavior) (**Figure S1G**). As a result, 97% of
213 possible TSS promoter variants were covered in each library on average (**Table S3**).

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

215 To ask how our libraries recapitulated known TSS efficiency measurements, we first
216 examined core sequences in our library matching the *SNR14* TSS and its variants
217 examined by Kuehner and Brow [22] for TSS efficiency (**Figure 2A**). Our randomized
218 library contains the *SNR14* TSS sequence embedded in our *SNR37* context along with
219 all single substitution variants of this sequence, including the subset previously
220 examined at *SNR14.* We found that Pol II MASTER recapitulated the single base effects
221 on TSS efficiency previously observed while also indicating that single base changes
222 around a TSS can have large effects on TSS efficiency.

223 Examining all designed +1 TSS variants present in our library, we focused our analysis
224 on positions -8, -1 and +1, which *in vivo* genome-wide data suggested are important
225 determinants for TSS selection (**Figure 2B, Figure S2A**). We first determined the TSS
226 efficiencies of the designed +1 TSS for all promoter variants, dividing all variants into
227 groups defined by bases at positions -8, -1 and +1 relative to TSS. This allowed us to
228 examine the TSS efficiency distribution within each subgroup. Our results not only
229 recapitulated the importance of these three positions in our defined promoter context
230 but also demonstrated that surrounding positions have a considerable impact on TSS
231 efficiency (note the wide range of efficiencies within each TSS group, **Figure 2B**). First,
232 in our controlled context for TSSs at the designed +1 position, we found that $Y_{-1}R_{+1}$ was
233 essential for initiation above a minimal background relative to $R_{-1}Y_{+1}$. Even in the
234 presence of an A_{-8}, $R_{-1}Y_{+1}$ variants showed essentially no usage. Second, we quantified
235 the very large effect of a -8A on TSS efficiency (note that $A_{-8}Y_{-1}R_{+1}$ motifs were much
236 higher in efficiency in aggregate than non-$A_{-8}Y_{-1}R_{+1}$ TSSs), demonstrating that -8A
237 alters TSS efficiency. This result is in line with the observation that $A_{-8}C_{-1}A_{+1}$ motif-
238 containing TSSs have the highest aggregate TSS usage from genomic promoters and
239 appear to be the most efficient [21]. Third, among $Y_{-1}R_{+1}$ elements we found a clear
240 hierarchy of efficiency – $C_{-1}A_{+1} > C_{-1}G_{+1} > T_{-1}A_{+1} \approx T_{-1}G_{+1}$ – that was not apparent from
241 genomic promoter TSS usage or contexts, likely due to promoter sequence biases
242 within genomic promoters.

243 As noted above, a wide range of TSS efficiencies were observed within every -8/-1/+1
244 group, suggesting contributions of other sequence positions to initiation efficiency. To
245 determine if base identities within -7 to -2 positions have similar effects regardless of -8,
246 -1 and +1 identity, we rank ordered all individual TSSs for each $N_{-8}N_{-1}N_{+1}$ motif by the
247 efficiency of their $A_{-8}C_{-1}A_{+1}$ version (**Figure S2A**). This rank ordering by $A_{-8}C_{-1}A_{+1}$
248 efficiency was predictive of efficiency ranks for -7 to -2 sequences with different bases
249 at positions -8, -1 and +1. This observation indicates important contributions from
250 positions beyond positions -8, -1 and +1 and that positions might function independently
251 to determine TSS efficiency (examined below).

252 We set out to determine the contributions to Pol II TSS efficiency of individual bases at
253 each position across our randomized region relative to the designed +1 TSS. To do so,

254 we calculated the TSS efficiencies of examined TSS variants and divided them into
255 subgroups based on bases at individual positions relative to the designed TSS +1
256 position (**Figure S2B**). Comparison of TSS efficiencies across base subgroups
257 suggested significant individual base effects on TSS efficiency in aggregate at all
258 examined positions. The usage of other positions in our promoter outside of the
259 designed TSS +1 position (either within the randomized region or outside of it) allowed
260 us to examine contribution of bases in our randomized region to the efficiency of these
261 other TSSs (**Figure S1E, F, Figure 2C**). Our combined libraries allowed us to analyze
262 efficiencies of nearly a million TSS variants present within them (distribution of
263 efficiencies is shown in **Figure S2C**). In order to visualize sequence preferences, we
264 used the median initiation efficiency values of each base subgroup as indicators for
265 preference, using centered median values to calculate "relative efficiency" and
266 illustrated this preference in a sequence logo (**Figure 2D**). Datasets of designed +1
267 TSSs deriving from "AYR", "BYR", and "ARY" libraries allowed us to nearly
268 comprehensively study preference at positions -8 to +1 (**Figure 2E**, middle and bottom).
269 Additionally, 10-25% of total TSS usage among different libraries deriving from a TSS at
270 +4 (**Figure 2C**, brown TSS arrow and sequences) allowed study of positions -11 to -9
271 relative to this TSS (**Figure 2E**, top). As noted above, positions -8, -1 and +1 were
272 major determinants for TSS efficiency. Interestingly, position -9 showed a relatively
273 strong effect in our defined promoter context, which was not described in genome-wide
274 analyses. At positions -4 to -2, we observed modulation of initiation efficiency, where in
275 general C and/or G were preferred and T was less-preferred. This overall preference is
276 consistent with an observation where individually mutating Ts at positions -4 to -2
277 relative to -38 TSS of *ADH1* promoter to a C significantly increased usage of that TSS
278 [57]. Though preferences at positions -7 to -5 were statistically significant (**Figure S2B**),
279 contributions were much lower than other examined positions. Taken together, these
280 results indicate that positions -9 to -8 and -4 to +1 are two major clusters contributing to
281 TSS efficiency.

282 Experiments across species and the description of the canonical initiator element (Inr)
283 suggest some sequence contributions from downstream positions relative to the TSS
284 [10,43,44,57-61]. To determine the impact of sequences downstream of the TSS, we
285 examined motif enrichment of the most efficient -8 TSS variants, whose positions +1 to
286 +9 are located in the randomized region (**Figure 2C**, green TSS arrow and sequences).
287 We found an $A(/G)_{+2}G(/C)_{+3}G(/C)_{+4}$ enrichment for the top 10% most efficient TSS, but
288 not for the next 10% most efficient TSS (**Figure S2D**). These preferences are consistent
289 with early mutagenesis work in yeast, where A(+2) to C/T, G(+3) to T, or C(+4) to T
290 substitutions decreased the utilization of a particular TSS, but A(+2) to G or T(+5) to C
291 substitutions had minor effects [57].

292 Pairwise nucleotide-position dependencies have been observed for some processes, for
293 example in 5' splice sites [62-64]. To investigate potential higher order sequence
294 interactions *i.e.* coupling between positions contributing to TSS efficiency, we examined
295 all possible pairwise interactions among positions -11 to +1 (**Figure 2F, G, Figure S2E,**
296 **F**). "Coupling" would entail a base at one position determining the contribution or effect
297 of a base at another position. We found evidence for coupling at multiple positions with
298 the strongest coupling between -9 and -8 positions. Here, an A at one position

suppresses the preference of A at the other position indicative of positional epistasis at each position where an A at one position might diminish the impact of an A at the other (**Figure S2E**). Using this -9/-8 interaction as an example, **Figure 2F** shows how coupling was detected and visualized. We calculated "centered relative efficiency" values for each base and position and visualized them as a heatmap (**Figure 2F, G**). We found that evident interactions were mainly observed at neighboring positions (**Figure 2G**), especially within positions -9 to -7 and within positions -5 to -2. In addition to its interaction with position -9 described above, position -8 was also observed to interact with position -7 (**Figure S2F**). These observed interactions suggest that these positions 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 [21]. As we noted in our prior work, this result could be a consequence of Pol II GOF or LOF mutants shifting TSSs upstream or downstream, respectively, coupled with the uneven distributions of bases and TSS motifs within genomic promoters. Furthermore, TSS motifs with a -8A by definition have an increased probability of having a potential upstream TSS that can be used (the -8A could function as an upstream +1R TSS but might lack its own -8A). As a result, upstream-shifting mutants seem to decrease preference for -8A TSSs because they are shifting from these to use of the -8A as a +1A. 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 3, Figure S3**). We introduced variant libraries into two Pol II LOF mutants (*rpb1* F1086S and H1085Q) and two GOF mutants (*rpb1* E1103G and G1097D). Data for all mutants showed high coverage of libraries (**Figure S3A**), were highly reproducible (**Figure S3B-E**), and showed expected upstream or downstream shifts in TSS usage in aggregate across libraries (**Figure S3B**) with broad effects on efficiencies across sequences (**Figure 3, Figure S3F-I**). In our previous work [21] we speculated that Pol II mutants with decreased catalytic activity (LOF) would decrease initiation across all sequences; conversely, alleles with increased catalytic activity (GOF) would increase initiation across all sequences. We find that this is indeed the case (**Figure 3A**). In our controlled promoter context, we find that LOF mutants decreased efficiencies across all sequences, and GOF mutants increased efficiencies across all sequences, and that mutants did not show strong effects on -8 (as was observed for usage at genomic promoters [21]). Instead, we observe differences in effects on +1A vs +1G (**Figure 3A**), and these effects are apparent across all levels of innate TSS efficiency (**Figure 3B**). Both GOF and LOF mutants show stronger effects on +1A relative to +1G, consistent with changes to the Pol II active site altering activity for different initiating NTPs.

Pol II initiation for +1G TSSs is sensitive to changes in GTP levels

We hypothesized that the effects on initiating +1 NTP could be direct effects of Pol II active site mutants altering activity in a substrate-selective fashion. However, it is also

342 conceivable that NTP levels within cells might be altered in Pol II mutants. To
343 investigate if the observed selective effects of Pol II mutants initiating NTP could be due
344 to altered NTP levels, we measured total NTP levels in WT, F1086S (LOF), and
345 E1103G (GOF) strains (**Figure 4A**). E1103G showed similar NTP levels as WT while
346 F1086S increased NTP levels to different extents. While both ATP and GTP in F1086S
347 were higher than WT, we reasoned that initiation might be more sensitive to GTP levels
348 as GTP is less in excess of the apparent K_D of Pol II for NTPs than is ATP. If so, this
349 provides an explanation for decreased deficit of +1G TSSs relative to +1A TSSs for
350 F1086S; increased GTP level in F1086S may buffer decreases in efficiency for +1G
351 TSSs.

352 If initiation is sensitive to GTP concentration in vivo, we hypothesized that reduction in
353 GTP should result in selectively reduced efficiency for +1G TSSs relative to +1A TSSs.
354 Therefore, we examined if and how initiation of our promoter variant libraries and
355 genomic TSSs was affected by treatment of cells with Mycophenolic Acid (MPA), which
356 is known to deplete GTP through inhibition of yeast inosine monophosphate
357 dehydrogenase (IMPDH) activities encoded by *IMD3* and *IMD4* [65,66]. We treated our
358 mixed AYR and BYR libraries with MPA for 20 min prior to Gal induction, while EtOH
359 treatment alone (the solvent for MPA) was used as a control. MPA decreased GTP level
360 as expected but also increased ATP, CTP, and UTP levels (**Figure 4B**).

361 MPA showed effects on promoter variant libraries in aggregate (**Figure S4A**) with an
362 overall downstream shift in TSS usage relative to EtOH alone. Comparison of replicates
363 indicated high reproducibility of MPA treatment (**Figure S4A-B**). When examining
364 specific sequences, we observed a striking and selective decrease in efficiency for
365 essentially all +1G TSSs, while +1A TSSs were almost entirely unaffected (**Figure 4C**).
366 MPA was previously shown to depress efficiency of +1G TSSs at *IMD2* as part of its
367 regulation by GTP levels [67]. Our results extend this observation to all +1G TSSs in our
368 MASTER library.

369 **Pol II initiation efficiencies are altered genome-wide upon changes to NTP levels**

370 To determine if these striking effects were observed in promoter contexts beyond our
371 designed library context, we performed TSS-seq on genome-derived RNAs from the
372 same samples that were analyzed for MASTER libraries. Examination of genomic TSSs
373 revealed effects of MPA on TSS efficiencies across the genome (**Figure 4D**). We
374 observed a depression of efficiency on average across all TSSs in MPA vs. EtOH
375 (vehicle) treatment. Given that we observed strong effects at +1G for MASTER libraries,
376 we speculated that there might also be effects on TSSs with +2G or potentially beyond,
377 especially because initial bond formation requires interaction with two substrates, and
378 subsequent initial elongation will lack stabilization of the product RNA that occurs with a
379 full-length (~9 nt) RNA:DNA hybrid. Examination of positions +2 through +6 indicates
380 that G positions through +5 are correlated with decreased initiation efficiency (**Figure**
381 **4D, S4D**). Conversely, we observed that a +2C or +2U was correlated with increased
382 efficiency relative to +2A or +2G. These results are consistent with the observed
383 increased CTP and UTP levels upon MPA treatment. We hypothesized that A positions
384 show decreased overall efficiency because most transcripts will have at least one G

385 within the first 5-6 nucleotides. To examine TSSs more carefully for G effects, we
386 examined TSSs lacking G entirely for the first 6 positions (**Figure 4E**). TSSs lacking a G
387 in the first 6 positions were more efficient on average than all TSSs in the presence of
388 MPA explaining some of the overall suppression of all TSS efficiencies by MPA. We
389 examined TSSs lacking a G in the first 6 positions and observed reduced effects on
390 efficiency from MPA treatment relative to all TSSs, consistent with this idea.
391 Furthermore, this analysis confirmed C and U effects relative to A at the +2 position.

392 **Learned initiation preferences are predictive of TSS efficiencies at genomic**
393 **promoters**

394 To ask how sequence determinants identified here relate to natural promoters, we
395 compared our library-defined sequence efficiencies to TSS efficiencies observed at
396 genomic promoters (**Figure 5**). To limit potentially confounding factors for genomic
397 promoters, we focused on a single "median" TSS for each promoter in a defined set of
398 promoter windows. The median TSS is defined as the TSS representing the position of
399 50th percentile of reads within each promoter window [21], i.e. the TSS representing the
400 middle of the cumulative distribution function of promoter reads from upstream to
401 downstream. We found that Pol II sequence preference at positions around median
402 TSSs was mostly consistent with what we observed in our libraries (**Figure 2E, Figure**
403 **5A**). Efficient genomic TSSs appear enriched for A at positions -7 to -5. The A-richness
404 at positions between -10 to -3 and +5 to +10 has been noted in previous studies from
405 our lab [21] (**Figure 5B**) and others [1,16,34]. However, As at positions -7 to -5 appeared
406 neutral in our promoter libraries (**Figure 2E**). The observed A-richness in genome could
407 reflect selection *in vivo* for additional promoter properties, such as providing an easily
408 meltable DNA region, lower nucleosome occupancy or reflect a context dependent role
409 not reflected by our promoter libraries.

410 We find that the interaction between -9A and -8A discovered in our libraries is also
411 reflected in genomic TSSs (**Figure 5C**). We first grouped the median TSSs from the top
412 20% expressed promoters based on the base at position -9 or -8 and then examined
413 sequence enrichment at the other position. We observed that -9A decreased the
414 enrichment of -8A relative to that when position -9 is not A (**Figure 5C**, left). Moreover,
415 when A was absent from the position -8, much higher enrichment for -9A was observed
416 (**Figure 5C**, right). Together, consistent with -9/-8 interaction discovered in our
417 controlled context, in the genome the presence of an A at either position -9 or -8
418 suppresses enrichment of A at the other position within the group of most efficiently
419 used median TSSs. These results suggest that -9A may function in similar fashion as -
420 8A, but that -8A may be more effective and therefore has been evolutionarily favored
421 (see Discussion).

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

424 We have found that DNA sequences around the TSS not only additively but also
425 interactively contribute to TSS efficiency. To quantitatively identify key features
426 (sequences and interactions) for TSS efficiency, we compiled datasets deriving from all

427 libraries and predicted TSS efficiency from sequence information by logistic regression
428 coupled with a forward stepwise selection strategy (**Figure 6A-E, Figure S5A-C**). We
429 first compiled datasets generated from designed -8 to +2 and +4 TSSs deriving from
430 "AYR", "BYR", and "ARY" libraries (**Figure S5A**) and split data into training (80%) and
431 test (20%) sets. For each variant, its sequences at positions -11 to +9 were extracted as
432 potential predictors for TSS efficiency. We then used a forward stepwise strategy with a
433 5-fold Cross-Validation (CV) to select robust features (predictors). By evaluating model
434 performance with R^2, sequences at nine positions (positions -9 to -7 and -4 to +2) and
435 one interaction (-9/-8 interaction) were identified as robust features and selected for final
436 modeling (**Figure 6A**). It is worth noting that models with as few as three features –
437 sequences at positions -9, -8, -1 (or +1) – could explain 74.10% of TSS efficiency
438 variation.

439 The final model containing the most predictive features explained 91.60% of the
440 variance in TSS efficiency for WT test set (20% of total dataset) (**Figure 6B, Figure**
441 **S5B**). Principal Component Analysis (PCA) on variables of models trained with
442 individual replicates of WT and Pol II mutants shows large differences between WT and
443 Pol II mutants but very small differences between replicates (**Figure S5C**), indicating
444 that modeling captured features of different Pol II groups. We next asked whether the
445 features learned by modeling using compiled datasets were consistent with our previous
446 sequence preference analysis using selected representative datasets with the most
447 randomized bases. We centered additive variable values and visualized as a sequence
448 logo (**Figure 6C**). First, as expected, positions -1 and +1 were the major predictors,
449 however the influence of the -8 A did not appear as strong as in our previous preference
450 analysis. We suspected this might be because the -9/-8 interaction contribution was not
451 included. After adding the -9/-8 interaction term, we observed emergence of the position
452 -8 as an influential predictor (**Figure 6D, E**), which also emphasizes the contribution of
453 the -9/-8 interaction. Second, and importantly, modeling confirmed the +2A preference
454 observed in previous motif enrichment analysis using only the most efficient -8 TSS
455 variants (**Figure S2D**). The impact of position +2 is also evident when performing
456 Principal Component Analysis (PCA) to variables of models trained with WT or Pol II
457 mutant datasets (**Figure S5C**). The fact that sequences at position +2 are the top
458 contributing variables in 2nd principal component that distinguished G1097D and
459 E1103G is in agreement with differentially altered +2 sequence enrichment by Pol II
460 mutants (**Figure S3I**). These results suggest position +2 preference is altered by Pol II
461 activity changes, and that this position might work directly with the Pol II active site.

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

464 To evaluate the extent to which DNA sequence around a TSS contributes to TSS
465 efficiency at genomic promoters, we compared the difference between observed and
466 model predicted efficiencies for all positions within known promoter windows or within
467 specific subgroups of known promoters (**Figure 6F-H, Figure S5D**). As expected, we
468 found most promoter positions showed low or no observed efficiency and were over-
469 predicted by sequence alone (**Figure 6F, Figure S5D**). This is because TSSs need to
470 be specified by a core promoter followed by scanning over a limited distance. Therefore,

471 features beyond TSS sequence, such as distance from the site of PIC assembly,
472 determine initiation in the genome. We therefore extracted observed median TSSs as
473 representatives of TSSs in contexts supporting efficient initiation to ask how our
474 sequence-based predictor functioned on genomic TSSs (**Figure 6G**). We also
475 separated median TSSs by promoter classes based on Taf1 enrichment [68], a proxy for
476 the two main types of promoters in yeast, and by promoter expression levels (**Figure**
477 **6H**). We observed good prediction performance at a wide range of TSSs indicating
478 sequence determinants identified in our limited promoter context contribute to TSS
479 efficiency in genomic promoter contexts. We observed increased performance for more
480 highly expressed promoters (Pearson r increased from 0.37-0.54 to 0.46-0.64) and
481 Taf1-depleted promoters (Pearson r increased from 0.37-0.50 to 0.54-0.64) (**Figure**
482 **6H**). Conceivably, highly expressed promoters may have evolved TSSs at optimal
483 distances from core promoters, and therefore may be similarly sensitive to sequence
484 effects. Alternatively, the use in our libraries of *GAL1* UAS and *SNR37* core promoter
485 elements, where both *GAL1* and *SNR37* are highly-expressed, Taf1-depleted
486 promoters, may share sequence sensitivities for TSSs from related promoters.

487 Discussion

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

502 Combining results from this study and others, we suggest how TSS sequence
503 contribution for TSS selection works (**Figure 7**). We propose that two major sequence
504 position groups contribute to TSS selection: bases around the TSS and bases around
505 position -8. First, in promoter scanning, the TSS and adjacent bases interact with Pol II
506 active site, the 1st NTP or each other to facilitate stable binding of 1st NTP and
507 potentially the 2nd NTP to stimulate RNA synthesis. Sensitivity of initiation efficiency to
508 NTP levels for first through fifth transcription positions in vivo supports this idea. This
509 function of downstream positions would be in contrast to the concept of the initiator or
510 downstream elements functioning as part of the TFIID or PIC binding site, as has been
511 proposed for higher eukaryotes (see [69,70] and references therein). As the universal
512 initiating element, $Y_{-1}R_{+1}$ has been established to facilitate stable binding of the NTPs by
513 RNA polymerases via base stacking between R_{-1} from template DNA and the 1st purine
514 NTP [43,44]. Positions upstream of the TSS, such as positions -4 to -2, might contribute to

515 stabilize template DNA via base stacking or physical interaction with initiation factors
516 [44,71].

517 Our observation that Pol II mutants showed selective effects on base at position +1
518 (**Figure 3**) supports that position +1 is interpreted by the Pol II active site. Meanwhile,
519 the observation that Pol II initiation is sensitive to NTP pools (**Figure 4C, Figure S4A**)
520 further suggests a mechanism for cellular state to regulate initiation via alteration of
521 NTP levels for the first position and several subsequent transcript positions levels. For
522 GOF Pol II mutants, we propose that differential preference for ATP versus GTP is a
523 direct consequence of Pol II active site changes. Pol II has been observed to bypass
524 cyclobutane pyrimidine dimers (CPD) in vitro through addition of an untemplated A
525 across from CPD lesions and a GOF polymerase appears to promote this activity,
526 suggestion a potential increase in selectivity for ATP in some situations [72]. In contrast to
527 effects observed for GOF mutants we proposed that tested LOF mutants may have
528 selective effects on TSS efficiency for +1A versus +1G sites due to indirect effects on
529 NTP levels in cells. Such defects might result from altered synthesis of nucleotide
530 synthesis-related genes, a number of which are themselves sensitive to Pol II activity
531 [23,67,73-76]. We observed the LOF mutant F1086S indeed alters all NTP levels (**Figure**
532 **4A**). Specifically, the LOF mutant increased GTP level, which is consistent with previous
533 observations of GTP-related genes, *IMD2*, *IMD3* and *IMD4*. In Pol II LOF mutants, *IMD2*
534 is constitutively active [23,74], and overall expression at the mRNA level for *IMD* genes is
535 increased [23].

536 While it is unsurprising that levels of NTP substrates would alter initiation efficiency, it is
537 striking that differences can be observed genome wide due to a modest reduction or
538 increase in GTP, CTP, or UTP levels. These changes suggest that environmental
539 control of initiation may be used more broadly than anticipated and that distinct
540 transcripts will be differentially affected upon perturbation of NTP levels. Direct sensing
541 and regulation of promoter activity by NTP levels has been proposed for the *IMD2*
542 promoter in yeast [67,77]. At this promoter, under GTP replete conditions upstream +1G
543 TSSs are used to generate a non-productive set of transcripts that terminate before
544 transcription of the *IMD2* reading frame. Upon GTP limitation (or catalytic defects in Pol
545 II), productive initiation is observed at a downstream +1A TSS. Our results here argue
546 that nucleotide sequence content is likely sufficient to confer the observed regulation as
547 we find that most +1G TSSs in the genome are reduced for initiation efficiency upon
548 reduction in GTP levels. For select TSSs from *IMD2* that they tested, the Brow lab had
549 previously found that Gs at the first or second transcript positions could confer MPA
550 sensitivity to a TSS [67]. Here, we find this effect genome-wide and that it extends to
551 transcript +5 position for GTP and also find evidence for effects of UTP and CTP levels.

552 The positions +2 to +4 downstream of TSS might contribute to establish initial RNA and
553 NTP stability in Pol II active site as well (**Figure S2D**). We observed Pol II mutant
554 effects on those preferences (**Figure S3I, Figure S5C**), suggesting these positions
555 might function via directly interacting with the Pol II active site, as observed in
556 transcription structures in other species [43,44]. First, the T7-like single-subunit RNAP
557 family shows a base-specific interaction between +2 NTP and a residue in the middle of
558 the O helix, which was suggested to enhance formation of the first phosphodiester bond

559 [43]. Second, a structure of *de novo* transcription initiation complex in bacterial RNA
560 Polymerase showed multiple interactions between the 2nd NTP and its β' and β
561 subunits, whose eukaryotic counterparts are Rpb1 and Rpb2 [44]. Alternatively, the A/G-
562 rich $A(/G)_{+1}A(/G)_{+2}G_{+3}G_{+4}$ motif might be related to translocation. A study from the
563 Landick lab showed that when A/G comprised the RNA 3' end, the RNAP active site
564 favored the post-translocated state [78]. If A/G similarly affects translocation state within
565 first four bases in Pol II initiation, synthesis of the first few bases might be promoted.
566 Our results detecting sensitivity in initiation efficiency based on levels of NTPs specified
567 at positions downstream from +1 (**Figure 4, S4**) suggest potential abortive initiation
568 under conditions that slow phosphodiester bond formation at positions early in transcript
569 synthesis. Abortive initiation by yeast Pol II in vitro under NTP replete conditions is very
570 low in specific artificial initiation conditions [79]. However, these conditions did not involve
571 initiation by a scanning PIC with all GTFs or reduced NTP levels. Revisiting the
572 biochemical properties of Pol II during bona fide initiation by promoter scanning could
573 be valuable and our studies make specific predictions that can be tested.

574 The GTF TFIIB has been specifically implicated in the preference for bases near
575 position -8, where the TFIIB B-reader domain has been observed to directly interact in a
576 structure of a yeast Pol II-TFIIB complex [80]. Here, it is attractive to envision TFIIB
577 functioning as an anchor point to pause the scanning process to promote Pol II initiation
578 at a fixed distance downstream. Several observations support this proposed function.
579 First, we detected sequence interaction between positions -8 and -7 (**Figure 2G, Figure**
580 **S2F**). This is in line with the direct contact of -7T and -8T on the template strand and
581 TFIIB B-reader R64 and D69 observed in Pol II-TFIIB complex structure [80], which has
582 been proposed to hold TSSs in Pol II active site longer during scanning. Second, we
583 observed a strong interaction between positions -9 and -8, where the presence of an A
584 at either position suppressed the preference of A at the other position (**Figure 2F,**
585 **Figure S2E**). This -9/-8 interaction was also evident when examining genomic median
586 TSSs (**Figure 5C**). Taken together, we speculate that Ts around position -8 or -9 on the
587 template strand and TFIIB may pause the scanning process to facilitate the usage of
588 TSSs positioned 8 to 9 bases downstream. Moreover, we have shown that Pol II
589 catalytic mutants alter TSS efficiencies across all TSS sequences, without showing
590 alteration in preference for -8A (**Figure 3, Figure S3F**), suggesting that Pol II catalytic
591 activity is not responsible for -8 preference.

592 Whether or how DNA sequence surrounding TSSs is involved in other promoter
593 properties is another question. A bases at positions -7 to -5 were observed to be neutral
594 in our promoter libraries (**Figure 2E**), in contrast to A-enrichment observed for highly
595 expressed and focused genomic TSSs (**Figure 5A, B**) [1,16,21,34,45]. We speculate that
596 observed A-richness around TSS functions through other evolved promoter properties.
597 <u>First</u>, observed A-richness between positions -10 to -3, together with T-richness at
598 further upstream core promoter region [45,46], provides an easily meltable region for DNA
599 unwinding, perhaps facilitating transcription initiation in specific contexts. <u>Second</u>,
600 higher A/T content may function to lower nucleosome occupancy [81], because
601 appropriately periodic G/C dinucleotides promote nucleosome occupancy [82-85].
602 However, the base composition switch in highly expressed promoters from T- to A-
603 preponderance [21,45,86] indicates A-richness may have other roles depending on the

604 characteristics of the sequence itself. Third, A-richness may be left over from the
605 evolution of promoter scanning. A recent study of transcription initiation mechanism
606 investigated 12 yeast species and proposed that during evolution a A-rich region
607 upstream of TSS appeared first, then the specific -8A preference occurred [47]. Therefore,
608 the observed A preference at upstream positions of highly used and focused TSSs may
609 be leftovers of A-enrichment in those promoters during scanning evolution in addition to
610 promoter roles beyond TSS selection per se.

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

623

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 [21,23,73,87]. 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. All plasmids and yeast strains are available upon request. Yeast media are following standard protocols [88]. 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 [88] with minor alterations as described in [23]: 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) [89] 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 [90]. For Pol II mutant libraries, 2 µg plasmid pool per reaction was electroporated into Pol II mutant strains following yeast electroporation transformation protocol described in [91], 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

667 x 10^7 cells/ml, as determined by cell counting. Subsequently, galactose (Amresco,
668 0637) was added for three hours (4% final concentration) to induce library expression.
669 For MPA treated WT libraries, MPA (in 100% ethanol; Sigma-Aldrich, M3536) was
670 added for 20 min (20 µg/ml final concentration) prior to the three-hour galactose
671 induction, with same volume of 100% ethanol as solvent control. 50 ml and 5 ml culture
672 aliquots, for RNA and DNA extraction respectively, were harvested and then cell pellets
673 were stored at -80 °C for further processing as described below.

674 **Generation of DNA amplicons for DNA-seq**

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

687 **Sample preparation for TSS-seq**

688 Total RNA was extracted by a phenol-chloroform method [92], followed by RNA
689 purification (RNeasy Mini kit, QIAGEN, 74104) with on-column DNase digestion
690 (RNase-Free DNase Set, QIAGEN, 79254) to remove DNA. TSS-seq was done using
691 procedures described in [93]. To prepare RNAs for the cDNA library construction, samples
692 were sequentially treated with Terminator 5'-Phosphate-Dependent Exonuclease
693 (Lucigen), Quick CIP (calf-intestine alkaline phosphatase, NEB) and Cap-Clip™ Acid
694 Pyrophosphatase (CellScript) to remove 5' monophosphate RNA and convert 5'
695 triphosphate or capped RNAs to 5' monophosphate RNAs. Next, RNA prepared with
696 enzymatic treatments was ligated to the 5'-adapter (s1206-N15, 5'-
697 GUUCAGAGUUCUACAGUCCGACGAUCNNNNNNNNNNNNNNN-3') that contains
698 Illumina adapter sequence and a 15 nt randomized 3'-end to reduce ligation bias and
699 serve as a Unique Molecular Identifier (UMI). Next, cDNA was constructed by reverse
700 transcription using RT primer CKO2191-s128A (5'-
701 CCTTGGCACCCGAGAATTCCAAGTGAATAATTCTTCACCTTTA-3') followed by
702 emulsion PCR amplification for 20-22 cycles using Illumina PCR primers (RP1 and
703 RPI3-30). Final DNA was gel size selected for 180-250 bp lengths and sequenced by
704 Illumina NextSeq 500 (150 SE) or NovaSeq 6000 (200 SE) (**Table S2**) using custom
705 primer s1115 (5'-CTACACGTTCAGAGTTCTACAGTCCGACGATC-3') to avoid
706 potentially confounding effects of misannealing of the default pooled Illumina
707 sequencing primers to the two randomized sequence regions.

708 **Primer extension assay**

709 Primer extension assays were performed on the same batch of total RNA extracted for
710 TSS-seq as described in [94] with modifications described in [23]. For each reaction, 30 µg
711 total RNA was used. An RNA sample without library transformed was used as “no GFP”
712 control. A sample containing same amount of nuclease-free water was used as “no
713 RNA” control. A primer (CKO2191) complementary to the 6th to 27th bases of *GFP* ORF,
714 which is the same annealing region for reverse transcription of TSS-seq sample
715 preparation, was labeled with ^{32}P γ-ATP (PerkinElmer, BLU502Z250UC) and T4
716 polynucleotide kinase (Thermo Scientific, EK0031). M-MuLV Reverse Transcriptase
717 (NEB, M0253L), RNase inhibitor (NEB, M0307L), dNTPs (GE) and DTT were added to
718 mix of RNA and labelled primer for reverse transcription reaction. Before loading to
719 sequencing gel, RNase A (Thermo Scientific, EN0531) was added to remove RNA. The
720 products were analyzed by 8% acrylamide/bis-acrylamide (19:1 ratio, Bio-Rad,
721 1610145) gel containing 1x TBE and 7M Urea. Primer extension gels were visualized by
722 a Molecular Imager PharosFX™ Plus System (Bio-Rad) and quantified by Image Lab
723 (5.2).

724 **Determination of NTP levels**

725 For each genotype or treatment, six biological replicates were performed. Cells from
726 saturated overnight cultures were used to inoculate fresh SC medium containing 2%
727 raffinose at 0.25 x 10^7 cells/ml and grown to a density of 0.5-0.8 x 10^7 cells/ml, as
728 determined by cell counting. For each WT replicate, two 15 ml cultures were aliquoted
729 and treated with 20 ug/ml MPA or 100% ethanol for 20 min, respectively. Subsequently,
730 three-hour 4% galactose treatment was performed for all samples. About 1 x 10^7 cells
731 were harvested, and then cell pellets were snap frozen using liquid nitrogen and
732 immediately stored at -80 °C for further processing as described below.

733 Metabolic quenching and nucleotide phosphate metabolite pool extraction were
734 performed by adding 1 ml ice-cold 2:2:1 acetonitrile:methanol:water with 0.1% formic
735 acid with 1x MS-Stop phosphatase inhibitor (Sigma-Aldrich). Samples were spiked with
736 10 µl 100 mM mix of adenosine-$^{15}N_5$-5′-triphosphate, guanosine-$^{15}N_5$-5′-triphosphate,
737 thymidine-$^{13}C_{10}$,$^{15}N_2$-5′-triphosphate, cytosine-$^{2}H_{14}$-5′-triphosphate, and uridine-$^{15}N_2$-5′-
738 triphosphate (Sigma-Aldrich). Samples were homogenized via sonification at room
739 temperature, and the supernatant was then cleared of protein by centrifugation at
740 16,000 xg. The protein pellet was resuspended in RIPA buffer for protein normalization.
741 The cleared supernatant was dried down under nitrogen gas and resuspended in 100 µl
742 7.5 mM ammonium acetate / 0.05% ammonium hydroxide. 2µl of cleared supernatant
743 was subjected to online LC-MS analysis. Purified adenosine-5′-triphosphate, guanosine-
744 5′-triphosphate, thymidine-5′-triphosphate, cytosine-5′-triphosphate, and uridine-5′-
745 triphosphate (Sigma-Aldrich) were serially diluted from 50 pmol/µl to 0.39 pmol/µl to
746 generate calibration curves.

747 Analyses were performed by untargeted LC-HRMS. Briefly, samples were injected via a
748 Thermo Vanquish UHPLC and separated over a reversed phase Phenomenex Kinetix-
749 Polar C18 column (2.1 x 100 mm, 1.7 µm particle size) maintained at 55°C. For the
750 22.5-minute LC gradient, the mobile phase consisted of the following: solvent A
751 (water/7.5 mM ammonium acetate/0.05% ammonium hydroxide) and solvent B

752 (acetonitrile/0.05% ammonium hydroxide). The gradient was the following: 0-.1 min 2%
753 B, increase to 70% B over 12 minutes, increase to 98% B over 0.1 min, hold at 98% B
754 for 5 minutes, re-equilibrate at 2% B for five minutes. The Thermo IDX tribrid mass
755 spectrometer was operated in positive ion mode, scanning in $ddMS^2$ mode (2 µscans)
756 from 200 to 800 m/z at 120,000 resolution with an AGC target of 2e5 for full scan, 2e4
757 for ms^2 scans using HCD fragmentation at stepped 15,35,50 collision energies. Source
758 ionization setting was 3.0 kV spray voltage for positive mode. Source gas parameters
759 were 35 sheath gas, 12 auxiliary gas at 320°C, and 8 sweep gas. Calibration was
760 performed prior to analysis using the Pierce™ FlexMix Ion Calibration Solutions
761 (Thermo Fisher Scientific). Integrated peak areas were then extracted manually using
762 Quan Browser (Thermo Fisher Xcalibur ver. 2.7). Calibration curves using purified
763 standards were then used to convert peak area ratios to concentration.

764 **Computational analyses**

765 Data and statistical analyses were performed in Python (3.8.5) and R (4.0.0)
766 environments. Additional packages usages are reported throughout the methods
767 description. Source code is provided at https://github.com/Kaplan-Lab-
768 Pitt/PolII_MASTER-TSS_sequence. Raw sequencing data have been deposited on the
769 NCBI SRA (Sequence Read Archive) under the BioProject accession number
770 PRJNA766624. Processed data have been deposited on the GEO (Gene Expression
771 Omnibus) under the accession number GSE185290. Visualizations were compiled in
772 Adobe Illustrator 2021.

773 ***DNA-seq analysis.*** High-throughput sequencing of template DNA amplicon was used
774 to assign each 9 nt randomized TSS sequence to a corresponding 24 nt barcode. First,
775 paired-end reads were merged using PEAR (0.9.11) [95]. Next, we considered only those
776 reads that contained a perfect match to three sequence regions common to all variants:
777 27 nt sequence upstream of the TSS region, 24 nt sequence between TSS region and
778 barcode, and 27 nt sequence downstream of barcode (5'-
779 TTCAAATTTTTCTTTTGATTTTTTTTCNNNNNNNNNACATTTTCAAAAGGCTAACATC
780 AGNNNNANNNNCNNNNTNNNNGNNNNATGTCTAAAGGTGAAGAATTATTCACT-3',
781 randomized TSS and barcode regions are underlined). From these reads, 9 nt TSS
782 region and 24 nt barcode were extracted, followed by individually error correction using
783 UMI-tools (1.0.0) [96]. Next, for barcodes linking to multiple TSS variants, only barcodes
784 for which >= 90% of the sequencing reads containing a specified barcode also
785 contained a shared, exact 9 nt TSS region were kept. To generate a master pool of
786 TSS-barcode linkages for all TSS-seq samples, for each library ("AYR", "BYR", "ARY"),
787 TSS-barcode linkages that existed in at least two out of four samples (one *E.coli* sample
788 plus three WT yeast replicates) and in which >= 5 reads existed were kept and pooled.
789 Two types of processed data are available in GEO database, with accession numbers
790 listed in (**Table S2**): tables containing TSS-barcode linkages and corresponding DNA-
791 seq read counts for each sample, tables of the master pool containing kept TSS-
792 barcode linkages and corresponding DNA-seq read count in all related samples.

793 ***TSS-seq analysis for libraries.*** High-throughput sequencing of RNA samples was
794 used to link RNA products to barcodes, therefore assign TSS usage to corresponding

795 DNA templates. For TSS identification and subsequent preference analysis, we
796 considered only those reads that contained a perfect match to a 27 nt sequence region
797 downstream of the barcode, as well as expected length of 5'-end: 5'-[15 nt 5'-UMI]-[>1
798 nt upstream of barcode region, designated as "RNA 5'-end"]-[24 nt barcode]-[the first 27
799 nt of GFP ORF, ATGTCTAAAGGTGAAGAATTATTCACT]-3'. Those reads carrying one
800 or more mismatches from expected sequences were used for analysis of relative
801 expression of promoters. Next, 15 nt 5'-UMIs, "RNA 5'-end" with varying length, and 24
802 nt barcode were extracted and individually corrected by UMI-tools (1.0.0). Deduplication
803 was performed based on 5'-UMIs, meaning reads contained a shared UMI-"RNA 5'-
804 end"-barcode linkage were counted as one deduplicated read for further analysis. Next,
805 the identity of the 24 nt barcode was used to determine the template sequences of
806 randomized TSS region. Then, reads with "RNA 5'-end" sequence perfectly matched to
807 corresponding template sequence were kept and used for downstream analysis. A TSS-
808 seq count table containing TSS usage distribution of each TSS promoter variant was
809 generated. In the count table, each row represents one TSS promoter variant, and each
810 column represents one position between positions -68 to +25 relative to "designed" +1
811 TSS. The number in each cell represents TSS-seq reads generated from a particular
812 position, with perfectly match to DNA template. After investigating reproducibility
813 (**Figure S3B-C, S4A-C**), count tables generated from three biological replicates were
814 merged into one by aggregating read counts at each position. Promoter variants with >=
815 5 TSS-seq reads in each replicate and whose Coefficient of Variation (CV) of TSS-seq
816 reads is less than 0.5 were kept. Two types of processed data are available in GEO
817 database, with accession numbers listed in **Table S2**: tables containing "RNA 5'-end"-
818 Barcode linkages and corresponding deduplicated TSS-seq read counts for individual
819 sample, TSS-seq count tables for individual samples and after aggregating replicates.
820 As an example, a TSS-seq count table including positions -10 to +25 relative to
821 "designed" +1 TSS after aggregating replicates of WT libraries is shown as **Table S4**.

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

826 ***Relative expression analysis for libraries.*** The relative expression from each library
827 promoter variant was defined as the ratio of normalized TSS-seq reads generated from
828 a particular promoter variant to normalized DNA-seq reads containing the variant.

829 ***Sequence preference analysis.*** For sequence preference at each position, all TSS
830 variants were subgrouped based on the bases at a particular position. TSS efficiencies
831 of TSS variants were visualized as scatter plots using GraphPad Prism 9. Kruskal-
832 Wallis with Dunn's test was performed to test sequence preference in GraphPad Prism
833 9. Next, TSS efficiency medians of each subgroup were calculated and centered to
834 calculate "relative efficiency" at each position. The relative efficiencies were visualized
835 as sequence logos using Logomaker (0.8) [97]. In motif enrichment analysis, surrounding
836 sequences relative to examined TSSs were extracted and visualized as sequence logos
837 using WebLogo 3 [98]. Heatmaps, scatter plots and density plots for comparing Pol II WT

838 and mutants were generated by Morpheus
839 (https://software.broadinstitute.org/morpheus) or ggplot2 (3.3.3) R package.

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

849 ***TSS-seq analysis for genomic TSSs.*** Genomic TSS-seq datasets from our lab's
850 previous study were used for comparison of model to in vivo TSS usage [27]. Quality
851 control, read trimming and mapping were performed as described in [21,27] to generate a
852 TSS count table that contains TSS-seq reads at each individual position within known
853 promoter windows ("median" TSS, 250 nt upstream and 150 nt downstream from
854 median TSS position). TSS efficiency calculation and subsequent sequence preference
855 analyses were performed as that for Pol II MASTER libraries. EtOH and MPA treatment
856 for genomic TSSs are described in Results and these libraries were analyzed as TSS-
857 seq was above. TSSs within a promoter were analyzed by their positions within the
858 cumulative distribution of usage from upstream to downstream, with 50% of the
859 distribution representing the median TSS position. To limit potential effects of marginal
860 TSSs, analyses of MPA treatment were limited to TSSs showing ≥2% efficiency in the
861 EtOH condition within the 25%-75% of the TSS distribution.

862 ***Prediction of TSS efficiency.*** To prepare datasets for modeling, positions of designed
863 -8 to +2 and +4 TSSs of each promoter variant that have valid TSS efficiency were
864 compiled as sequence variants. For each TSS variant, sequences at -11 to +9 positions
865 relative to TSS, together with corresponding TSS efficiency, were extracted. 80% of
866 dataset were randomly partitioned as training set and the rest 20% as testing set. To
867 select robust features, a forward stepwise strategy with a 5-fold Cross-Validation (CV)
868 was employed in two major stages, for additive terms and for interactions. Starting with
869 no variable in the model, logistic regression models with one additional variable (the
870 sequence at a particular position) were trained to predict TSS efficiency on training set
871 by train() of caret (6.0.86) R package [99], with a 5-fold CV. The R^2, representing the
872 proportion of variance explained, was calculated to indicate the performance of each
873 model. The variable that provides the highest increased R^2 for model was added into
874 the model for next round of variable selection. This process was repeated until the
875 increased R^2 is less than 0.01. After identifying the most influential additive variables,
876 same process was repeated for investigating robust interactions between selected
877 additive variables. Next, a final model with selected robust features. including additive
878 variables and interactions, was constructed on entire training set using glm() and
879 investigated on testing set. Comparison between predicted and measured efficiencies
880 was visualized as scatter plots using LSD (4.1.0) R package. Model parameters were
881 extracted and used to further calculation. Visualizations were done in Logomaker (0.8)

882 and Seaborn (0.11.0) in Python. Principal Component Analysis (PCA) was performed
883 using prcomp() in R.

884

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.

Supplementary Table 4. TSS-seq count table of WT libraries.

Acknowledgements

The authors thank Kaplan lab members for helpful comments on the manuscript. We are deeply grateful to Chenxi Qiu for discussions and comments on this project. We acknowledge Justin Kinney (Cold Spring Harbor Laboratory) and Shuoran Li (Statistical Consulting Center at University of Pittsburgh) for discussions on modeling. We thank Charles D. Johnson, Richard Metz (Texas A&M AgriLife Genomics and Bioinformatics Service), Andrew Hillhouse (Texas A&M Institute for Genome Sciences & Society), William A MacDonald, Rania Elbakri (the University of Pittsburgh Health Sciences Sequencing Core at UPMC Children's Hospital of Pittsburgh), Yinghong Pan (the UPMC Genome Center), Dibyendu Kumar (the Waksman Genomics Core Facility at Rutgers University), and Liz Freeman (Illumina) for discussions and advice regarding deep sequencing strategies. We thank Steven J. Mullett and Stacy Gelhaus Wendell (Metabolomics and Lipidomics Core, NIHS10OD023402) for performing NTP measurements.

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. S-H.Z. analyzed data and discussed analysis. B.E.N. provided funding and methodology of TSS-seq, and revised 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 and R35GM144116 to C.D.K. and R35GM118059 to B.E.N. This research was supported in part by the University of Pittsburgh Center for Research Computing, RRID:SCR_022735, through the resources provided. Specifically, this work used the HTC cluster, which is supported by NIH award number S10OD028483.

Data availability

Raw sequencing data generated in this study are available in the NCBI BioProject, under the accession number PRJNA766624. Processed data are available in GEO, under the accession number GSE185290.

Code availability

Code for analyses in this study is provided at https://github.com/Kaplan-Lab-Pitt/PolII_MASTER-TSS_sequence.

927 # References

928 1. Zhang, Z. & Dietrich, F.S. Mapping of transcription start sites in Saccharomyces
929 cerevisiae using 5' SAGE. *Nucleic Acids Res* **33**, 2838-51 (2005).
930 2. Park, D., Morris, A.R., Battenhouse, A. & Iyer, V.R. Simultaneous mapping of
931 transcript ends at single-nucleotide resolution and identification of widespread
932 promoter-associated non-coding RNA governed by TATA elements. *Nucleic*
933 *Acids Res* **42**, 3736-49 (2014).
934 3. Pelechano, V., Wei, W. & Steinmetz, L.M. Extensive transcriptional heterogeneity
935 revealed by isoform profiling. *Nature* **497**, 127-31 (2013).
936 4. Chia, M. et al. High-resolution analysis of cell-state transitions in yeast suggests
937 widespread transcriptional tuning by alternative starts. *Genome Biol* **22**, 34
938 (2021).
939 5. Nepal, C. et al. Dynamic regulation of the transcription initiation landscape at
940 single nucleotide resolution during vertebrate embryogenesis. *Genome Res* **23**,
941 1938-50 (2013).
942 6. Consortium, F. et al. A promoter-level mammalian expression atlas. *Nature* **507**,
943 462-70 (2014).
944 7. Yamashita, R. et al. Genome-wide characterization of transcriptional start sites in
945 humans by integrative transcriptome analysis. *Genome Res* **21**, 775-89 (2011).
946 8. Carninci, P. et al. The transcriptional landscape of the mammalian genome.
947 *Science* **309**, 1559-63 (2005).
948 9. Hoskins, R.A. et al. Genome-wide analysis of promoter architecture in Drosophila
949 melanogaster. *Genome Res* **21**, 182-92 (2011).
950 10. Zheng, H. et al. Global identification of transcription start sites in the genome of
951 Apis mellifera using 5'LongSAGE. *J Exp Zool B Mol Dev Evol* **316**, 500-14
952 (2011).
953 11. Chen, R.A. et al. The landscape of RNA polymerase II transcription initiation in C.
954 elegans reveals promoter and enhancer architectures. *Genome Res* **23**, 1339-47
955 (2013).
956 12. Cheng, Z. et al. Pervasive, Coordinated Protein-Level Changes Driven by
957 Transcript Isoform Switching during Meiosis. *Cell* **172**, 910-923 e16 (2018).
958 13. Rojas-Duran, M.F. & Gilbert, W.V. Alternative transcription start site selection
959 leads to large differences in translation activity in yeast. *RNA* **18**, 2299-305
960 (2012).
961 14. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T.R. High-fidelity
962 promoter profiling reveals widespread alternative promoter usage and
963 transposon-driven developmental gene expression. *Genome Res* **23**, 169-80
964 (2013).
965 15. Zhang, P. et al. Relatively frequent switching of transcription start sites during
966 cerebellar development. *BMC Genomics* **18**, 461 (2017).
967 16. Lu, Z. & Lin, Z. Pervasive and dynamic transcription initiation in Saccharomyces
968 cerevisiae. *Genome Res* **29**, 1198-1210 (2019).
969 17. Demircioglu, D. et al. A Pan-cancer Transcriptome Analysis Reveals Pervasive
970 Regulation through Alternative Promoters. *Cell* **178**, 1465-1477 e17 (2019).

971 18. Thorsen, K. et al. Tumor-specific usage of alternative transcription start sites in
972 colorectal cancer identified by genome-wide exon array analysis. *BMC Genomics*
973 **12**, 505 (2011).
974 19. Boyd, M. et al. Characterization of the enhancer and promoter landscape of
975 inflammatory bowel disease from human colon biopsies. *Nat Commun* **9**, 1661
976 (2018).
977 20. Lis, C.G.a.J.T. DNA Melting on Yeast RNA Polymerase II Promoter. *Science*
978 **261**, 759-762 (1993).
979 21. Qiu, C. et al. Universal promoter scanning by Pol II during transcription initiation
980 in Saccharomyces cerevisiae. *Genome Biology* **21**(2020).
981 22. Kuehner, J.N. & Brow, D.A. Quantitative analysis of in vivo initiator selection by
982 yeast RNA polymerase II supports a scanning model. *J Biol Chem* **281**, 14119-28
983 (2006).
984 23. Kaplan, C.D., Jin, H., Zhang, I.L. & Belyanin, A. Dissection of Pol II trigger loop
985 function and Pol II activity-dependent control of start site selection in vivo. *PLoS*
986 *Genet* **8**, e1002627 (2012).
987 24. Miller, G. & Hahn, S. A DNA-tethered cleavage probe reveals the path for
988 promoter DNA in the yeast preinitiation complex. *Nat Struct Mol Biol* **13**, 603-10
989 (2006).
990 25. Fazal, F.M., Meng, C.A., Murakami, K., Kornberg, R.D. & Block, S.M. Real-time
991 observation of the initiation of RNA polymerase II transcription. *Nature* **525**, 274-
992 7 (2015).
993 26. Hampsey, M. Molecular genetics of the RNA polymerase II general
994 transcriptional machinery. *Microbiol Mol Biol Rev* **62**, 465-503 (1998).
995 27. Zhao, T. et al. Ssl2/TFIIH function in transcription start site scanning by RNA
996 Polymerase II in Saccharomyces cerevisiae. *Elife* **10**(2021).
997 28. Hahn S, H.E., Guarente L. Each of three "TATA elements" specifies a subset of
998 the transcription initiation sites at the CYC-1 promoter of Saccharomy
999 cescerevisiae. *Proceedings of the National Academy of Sciences* **82**, 8562-6
1000 (1985).
1001 29. Cortes, T. et al. Genome-wide mapping of transcriptional start sites defines an
1002 extensive leaderless transcriptome in Mycobacterium tuberculosis. *Cell Rep* **5**,
1003 1121-31 (2013).
1004 30. Bucher, P. Weight matrix descriptions of four eukaryotic RNA polymerase II
1005 promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol*
1006 **212**, 563-78 (1990).
1007 31. Smale, S.T. & Baltimore, D. The "initiator" as a transcription control element. *Cell*
1008 **57**, 103-13 (1989).
1009 32. Corden, J. et al. Promoter sequences of eukaryotic protein-coding genes.
1010 *Science* **209**, 1406-14 (1980).
1011 33. McNeil, J.B. & Smith, M. Saccharomyces cerevisiae CYC1 mRNA 5'-end
1012 positioning: analysis by in vitro mutagenesis, using synthetic duplexes with
1013 random mismatch base pairs. *Mol Cell Biol* **5**, 3545-51 (1985).
1014 34. Malabat, C., Feuerbach, F., Ma, L., Saveanu, C. & Jacquier, A. Quality control of
1015 transcription start site selection by nonsense-mediated-mRNA decay. *Elife*
1016 **4**(2015).

1017 35. Policastro, R.A., Raborn, R.T., Brendel, V.P. & Zentner, G.E. Simple and efficient
1018 profiling of transcription initiation and transcript levels with STRIPE-seq. *Genome*
1019 *Res* **30**, 910-923 (2020).
1020 36. Healy, A.M., Helser, T.L. & Zitomer, R.S. Sequences required for transcriptional
1021 initiation of the Saccharomyces cerevisiae CYC7 genes. *Mol Cell Biol* **7**, 3785-91
1022 (1987).
1023 37. Furter-Graves, E.M. & Hall, B.D. DNA sequence elements required for
1024 transcription initiation of the Schizosaccharomyces pombe ADH gene in
1025 Saccharomyces cerevisiae. *Mol Gen Genet* **223**, 407-16 (1990).
1026 38. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture
1027 and evolution. *Nat Genet* **38**, 626-35 (2006).
1028 39. Hashimoto, S. et al. 5'-end SAGE for the analysis of transcriptional start sites.
1029 *Nat Biotechnol* **22**, 1146-9 (2004).
1030 40. Suzuki, Y. et al. Diverse transcriptional initiation revealed by fine, large-scale
1031 mapping of mRNA start sites. *EMBO Rep* **2**, 388-93 (2001).
1032 41. Kim, D. et al. Comparative Analysis of Regulatory Elements between Escherichia
1033 coli and Klebsiella pneumoniae by Genome-Wide Transcription Start Site
1034 Profiling. *Plos Genetics* **8**(2012).
1035 42. Vvedenskaya, I.O. et al. Massively Systematic Transcript End Readout,
1036 "MASTER": Transcription Start Site Selection, Transcriptional Slippage, and
1037 Transcript Yields. *Mol Cell* **60**, 953-65 (2015).
1038 43. Gleghorn, M.L., Davydova, E.K., Basu, R., Rothman-Denes, L.B. & Murakami,
1039 K.S. X-ray crystal structures elucidate the nucleotidyl transfer reaction of
1040 transcript initiation using two nucleotides. *Proceedings of the National Academy*
1041 *of Sciences of the United States of America* **108**, 3566-3571 (2011).
1042 44. Basu, R.S. et al. Structural basis of transcription initiation by bacterial RNA
1043 polymerase holoenzyme. *J Biol Chem* **289**, 24549-59 (2014).
1044 45. Maicas, E. & Friesen, J.D. A sequence pattern that occurs at the transcription
1045 initiation region of yeast RNA polymerase II promoters. *Nucleic Acids Res* **18**,
1046 3387-93 (1990).
1047 46. Lubliner, S., Keren, L. & Segal, E. Sequence features of yeast and human core
1048 promoters that are predictive of maximal promoter activity. *Nucleic Acids Res* **41**,
1049 5569-81 (2013).
1050 47. Lu, Z. & Lin, Z. The origin and evolution of a distinct mechanism of transcription
1051 initiation in yeasts. *Genome Res* (2020).
1052 48. Dujon, B. The yeast genome project: what did we learn? *Trends Genet* **12**, 263-
1053 70 (1996).
1054 49. Lubliner, S. et al. Core promoter sequence in yeast is a major determinant of
1055 expression level. *Genome Research* **25**, 1008-1017 (2015).
1056 50. Blazeck, J., Garg, R., Reed, B. & Alper, H.S. Controlling promoter strength and
1057 regulation in Saccharomyces cerevisiae using synthetic hybrid promoters.
1058 *Biotechnol Bioeng* **109**, 2884-95 (2012).
1059 51. Dhillon, N. et al. Permutational analysis of Saccharomyces cerevisiae regulatory
1060 elements. *Synth Biol (Oxf)* **5**, ysaa007 (2020).

1061 52. Wang, H., Schilbach, S., Ninov, M., Urlaub, H. & Cramer, P. Structures of
1062 transcription preinitiation complex engaged with the +1 nucleosome. *Nat Struct*
1063 *Mol Biol* (2022).
1064 53. Vvedenskaya, I.O., Goldman, S.R. & Nickels, B.E. Analysis of Bacterial
1065 Transcription by "Massively Systematic Transcript End Readout," MASTER.
1066 *Methods Enzymol* **612**, 269-302 (2018).
1067 54. Vvedenskaya, I.O. et al. Interactions between RNA polymerase and the core
1068 recognition element are a determinant of transcription start site selection. *Proc*
1069 *Natl Acad Sci U S A* **113**, E2899-905 (2016).
1070 55. Winkelman, J.T. et al. Multiplexed protein-DNA cross-linking: Scrunching in
1071 transcription start site selection. *Science* **351**, 1090-3 (2016).
1072 56. Hochschild, A. Mastering Transcription: Multiplexed Analysis of Transcription
1073 Start Site Sequences. *Mol Cell* **60**, 829-31 (2015).
1074 57. Faitar, S.L., Brodie, S.A. & Ponticelli, A.S. Promoter-specific shifts in transcription
1075 initiation conferred by yeast TFIIB mutations are determined by the sequence in
1076 the immediate vicinity of the start sites. *Molecular and Cellular Biology* **21**, 4427-
1077 4440 (2001).
1078 58. Deshpande, A.P. & Patel, S.S. Mechanism of transcription initiation by the yeast
1079 mitochondrial RNA polymerase. *Biochim Biophys Acta* **1819**, 930-8 (2012).
1080 59. Javahery, R., Khachi, A., Lo, K., Zenzie-Gregory, B. & Smale, S.T. DNA
1081 sequence requirements for transcriptional initiator activity in mammalian cells.
1082 *Mol Cell Biol* **14**, 116-27 (1994).
1083 60. Arkhipova, I.R. Promoter elements in Drosophila melanogaster revealed by
1084 sequence analysis. *Genetics* **139**, 1359-69 (1995).
1085 61. Yarden, G., Elfakess, R., Gazit, K. & Dikstein, R. Characterization of sINR, a
1086 strict version of the Initiator core promoter element. *Nucleic Acids Res* **37**, 4234-
1087 46 (2009).
1088 62. Wong, M.S., Kinney, J.B. & Krainer, A.R. Quantitative Activity Profile and Context
1089 Dependence of All Human 5' Splice Sites. *Mol Cell* **71**, 1012-1026 e3 (2018).
1090 63. Roca, X. et al. Features of 5'-splice-site efficiency derived from disease-causing
1091 mutations and comparative genomics. *Genome Res* **18**, 77-87 (2008).
1092 64. Carmel, I., Tal, S., Vig, I. & Ast, G. Comparative analysis detects dependencies
1093 among the 5' splice-site positions. *RNA* **10**, 828-40 (2004).
1094 65. McPhillips, C.C., Hyle, J.W. & Reines, D. Detection of the mycophenolate-
1095 inhibited form of IMP dehydrogenase in vivo. *Proc Natl Acad Sci U S A* **101**,
1096 12171-6 (2004).
1097 66. Hyle, J.W., Shaw, R.J. & Reines, D. Functional distinctions between IMP
1098 dehydrogenase genes in providing mycophenolate resistance and guanine
1099 prototrophy to yeast. *J Biol Chem* **278**, 28470-8 (2003).
1100 67. Kuehner, J.N. & Brow, D.A. Regulation of a eukaryotic gene by GTP-dependent
1101 start site selection and transcription attenuation. *Mol Cell* **31**, 201-11 (2008).
1102 68. Rhee, H.S. & Pugh, B.F. Genome-wide structure and organization of eukaryotic
1103 pre-initiation complexes. *Nature* **483**, 295-301 (2012).
1104 69. Vo ngoc, L., Huang, C.Y., Cassidy, C.J., Medrano, C. & Kadonaga, J.T.
1105 Identification of the human DPR core promoter element using machine learning.
1106 *Nature* **585**, 459-463 (2020).

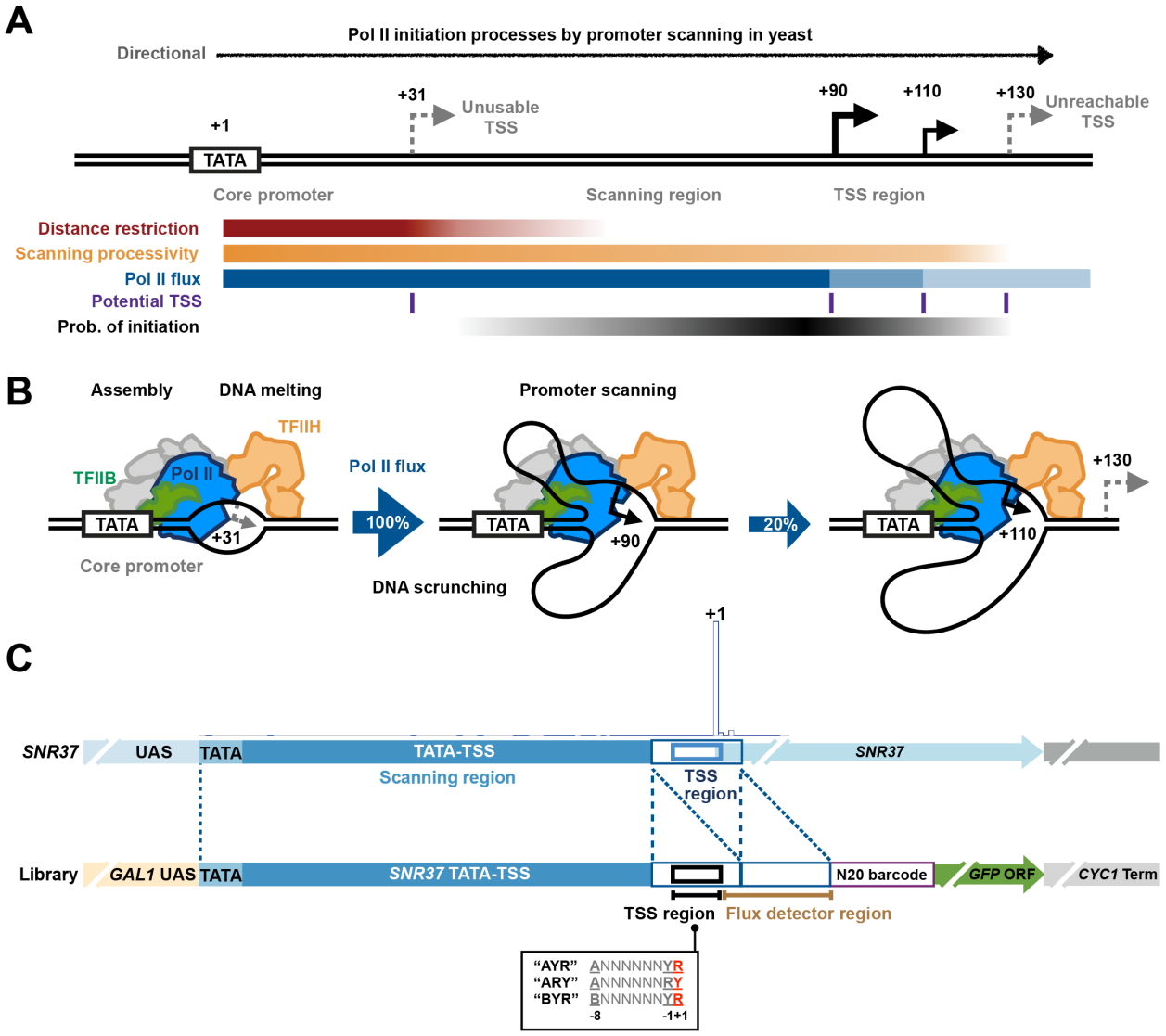
1107 70. Luse, D.S., Parida, M., Spector, B.M., Nilson, K.A. & Price, D.H. A unified view of
1108 the sequence and functional organization of the human RNA polymerase II
1109 promoter. *Nucleic Acids Res* **48**, 7767-7785 (2020).
1110 71. Zhang, Y. et al. Structural basis of transcription initiation. *Science* **338**, 1076-80
1111 (2012).
1112 72. Walmacq, C. et al. Mechanism of translesion transcription by RNA polymerase II
1113 and its role in cellular resistance to DNA damage. *Mol Cell* **46**, 18-29 (2012).
1114 73. Braberg, H. et al. From structure to systems: high-resolution, quantitative genetic
1115 analysis of RNA polymerase II. *Cell* **154**, 775-88 (2013).
1116 74. Malik, I., Qiu, C., Snavely, T. & Kaplan, C.D. Wide-ranging and unexpected
1117 consequences of altered Pol II catalytic activity in vivo. *Nucleic Acids Res* **45**,
1118 4431-4451 (2017).
1119 75. Kwapisz, M. et al. Mutations of RNA polymerase II activate key genes of the
1120 nucleoside triphosphate biosynthetic pathways. *EMBO J* **27**, 2411-21 (2008).
1121 76. Thiebaut, M. et al. Futile cycle of transcription initiation and termination
1122 modulates the response to nucleotide shortage in S. cerevisiae. *Mol Cell* **31**, 671-
1123 82 (2008).
1124 77. Steinmetz, E.J. et al. Genome-wide distribution of yeast RNA polymerase II and
1125 its control by Sen1 helicase. *Mol Cell* **24**, 735-746 (2006).
1126 78. Hein, P.P., Palangat, M. & Landick, R. RNA transcript 3'-proximal sequence
1127 affects translocation bias of RNA polymerase. *Biochemistry* **50**, 7002-14 (2011).
1128 79. Cabart, P., Jin, H., Li, L. & Kaplan, C.D. Activation and reactivation of the RNA
1129 polymerase II trigger loop for intrinsic RNA cleavage and catalysis. *Transcription*
1130 **5**, e28869 (2014).
1131 80. Sainsbury, S., Niesser, J. & Cramer, P. Structure and function of the initially
1132 transcribing RNA polymerase II-TFIIB complex. *Nature* **493**, 437-40 (2013).
1133 81. Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome
1134 organization. *Curr Opin Struct Biol* **19**, 65-71 (2009).
1135 82. Tillo, D. & Hughes, T.R. G+C content dominates intrinsic nucleosome occupancy.
1136 *BMC Bioinformatics* **10**, 442 (2009).
1137 83. Lee, W. et al. A high-resolution atlas of nucleosome occupancy in yeast. *Nat*
1138 *Genet* **39**, 1235-44 (2007).
1139 84. Peckham, H.E. et al. Nucleosome positioning signals in genomic DNA. *Genome*
1140 *Res* **17**, 1170-7 (2007).
1141 85. Segal, E. et al. A genomic code for nucleosome positioning. *Nature* **442**, 772-8
1142 (2006).
1143 86. Wu, R. & Li, H. Positioned and G/C-capped poly(dA:dT) tracts associate with the
1144 centers of nucleosome-free regions in yeast promoters. *Genome Res* **20**, 473-84
1145 (2010).
1146 87. Jin, H. & Kaplan, C.D. Relationships of RNA polymerase II genetic interactors to
1147 transcription start site usage defects and growth in Saccharomyces cerevisiae.
1148 *G3 (Bethesda)* **5**, 21-33 (2014).
1149 88. Amberg, D.C., Burke, D., Strathern, J.N., Burke, D. & Cold Spring Harbor
1150 Laboratory. *Methods in yeast genetics : a Cold Spring Harbor Laboratory course*
1151 *manual*, xvii, 230 p. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor,
1152 N.Y., 2005).

1153 89. Chee, M.K. & Haase, S.B. New and Redesigned pRS Plasmid Shuttle Vectors for
1154 Genetic Manipulation of Saccharomycescerevisiae. *G3 (Bethesda)* **2**, 515-26
1155 (2012).
1156 90. Gietz, R.D. & Schiestl, R.H. High-efficiency yeast transformation using the
1157 LiAc/SS carrier DNA/PEG method. *Nat Protoc* **2**, 31-4 (2007).
1158 91. Benatuil, L., Perez, J.M., Belk, J. & Hsieh, C.M. An improved yeast
1159 transformation method for the generation of very large human antibody libraries.
1160 *Protein Eng Des Sel* **23**, 155-9 (2010).
1161 92. Schmitt, M.E., Brown, T.A. & Trumpower, B.L. A rapid and simple method for
1162 preparation of RNA from Saccharomyces cerevisiae. *Nucleic Acids Res* **18**,
1163 3091-2 (1990).
1164 93. Vvedenskaya, I.O., Goldman, S.R. & Nickels, B.E. Preparation of cDNA libraries
1165 for high-throughput RNA sequencing analysis of RNA 5' ends. *Methods Mol Biol*
1166 **1276**, 211-28 (2015).
1167 94. Ranish, J.A. & Hahn, S. The Yeast General Transcription Factor Tfiia Is
1168 Composed of 2 Polypeptide Subunits. *Journal of Biological Chemistry* **266**,
1169 19320-19327 (1991).
1170 95. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate
1171 Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614-20 (2014).
1172 96. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in
1173 Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**,
1174 491-499 (2017).
1175 97. Tareen, A. & Kinney, J.B. Logomaker: beautiful sequence logos in Python.
1176 *Bioinformatics* **36**, 2272-2274 (2020).
1177 98. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence
1178 logo generator. *Genome Res* **14**, 1188-90 (2004).
1179 99. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of*
1180 *Statistical Software* **28**, 1-26 (2008).

1181

1182

1183 # 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"
CANNNNNRYACATTTTCAAAAGGCTAACAT
+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)

1184

1185 **Figure 1. A high-throughput system for studying transcription TSS selection. (A**
1186 **and B)** Pol II initiation in yeast proceeds by a promoter scanning mechanism. Yeast Pol
1187 II initiation usually occurs at multiple TSSs ranging from 40 to 120 bp downstream of a
1188 core promoter region comprising the PIC assembly position (the TATA-box for TATA-
1189 containing promoters). After the PIC assembles upstream (core promoter) scanning will
1190 proceed over a scanning region towards positions where TSS selection occurs (TSS
1191 region). In addition to sequence control of initiation through suitable TSS sequences,
1192 the probability of initiation across promoter positions is also controlled by multiple
1193 architectural features shown in **(A)**. These include the inhibition of initiation near a core
1194 promoter that diminishes as scanning proceeds ("distance restriction"), biochemical
1195 restrictions on how far scanning can proceed that increase as scanning proceeds
1196 ("scanning processivity"), and "Pol II flux", which represents the decrease in amount of
1197 scanning Pol II as scanning proceeds due to conversion of scanning Pol II to
1198 transcribing Pol II upon initiation. **(C)** Construction of promoter libraries that control TSS
1199 sequence context. Top panel shows schematic of the *SNR37* promoter and its TSS
1200 usage distribution based on TSS-seq [21]. *SNR37* has a focused and highly efficient TSS
1201 region. Bottom panel shows schematic of the Pol II MASTER libraries used in this study.
1202 A duplication of the *SNR37* TSS region was inserted before native TSS region, and the
1203 -8 to +1 positions relative to native *SNR37* +1 TSS (black box) were replaced by a 9 nt
1204 randomized region to provide almost all possible sequences. The second *SNR37* TSS
1205 region functions as a "Flux Detector" (FD) to capture Pol II flux that fails to initiate within
1206 the randomized region and allow determination of initiation efficiency within the
1207 randomized region. A barcode region (purple box) allows RNA products to be assigned
1208 to respective promoter variants. The *GAL1* UAS allows control of library expression, and
1209 the *GFP* ORF and *CYC1* terminator support termination and stabilize RNA products. **(D)**
1210 TSS usage distributions at TSS and FD regions for different promoter variant "AYR",
1211 "BYR", and "ARY" libraries are shown on the left. TSS usages from designed +1 TSS
1212 and positions upstream are in red and grey, respectively. TSS usage from the FD region
1213 is in brown. The definition of "TSS efficiency" and overall TSS efficiency calculations for
1214 the aggregate +1 TSS of all in different libraries are shown in the middle. Example
1215 correlation plots of TSS efficiency calculations for +1 TSSs from individual promoter
1216 variants in Pol II MASTER libraries between representative biological replicates are
1217 show on right. Pearson r and number of compared variants are shown.

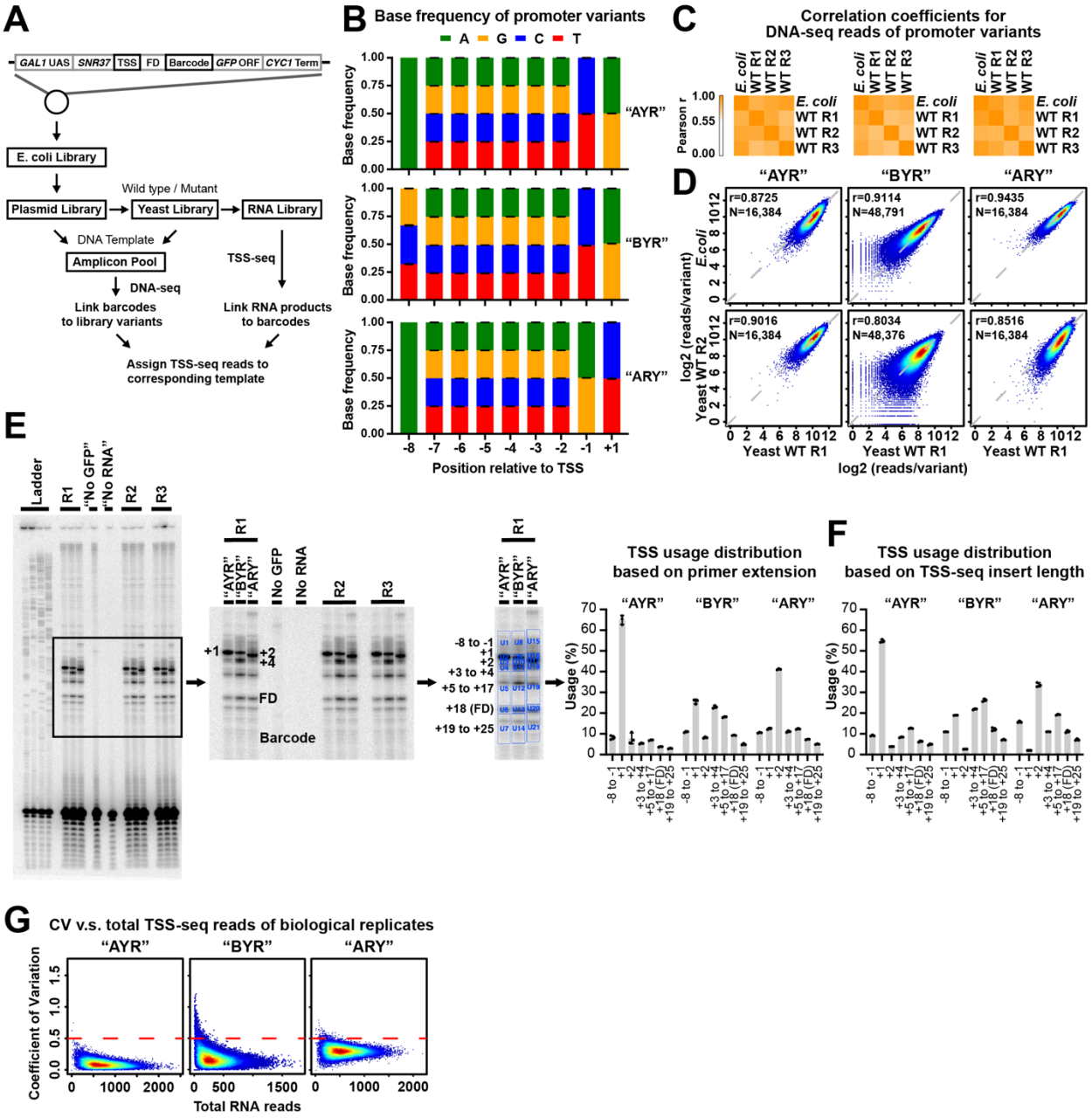


Figure S1. High level of reproducibility and coverage depth of library variants. (A) Schematic representation of experimental approach. Promoter libraries with almost all possible sequences within a 9 nt randomized region in a promoter context designed for specific functionalities were constructed on 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 the randomized region of promoter variants demonstrate unbiased synthesis of randomized regions. Bars are mean +/- standard deviation of the mean for promoter variants in WT and four

1229 Pol II mutants. **(C)** Heatmap illustrating hierarchical clustering of Pearson correlation
1230 coefficients of reads per promoter variant for libraries amplified in *E. coli* and three
1231 biological replicates of libraries transformed into yeast. **(D)** Example correlation plots of
1232 DNA reads count of promoter variants for *E. coli* and yeast WT biological replicates.
1233 Pearson r and number of compared variants are shown. **(E)** Bulk primer extension for
1234 RNA produced from promoter variant libraries and quantification for biological replicates
1235 transformed into WT yeast. “No GFP” control used an RNA sample without a library
1236 transformed. “No RNA” control used a sample of nuclease-free water. Dots represent
1237 three biological replicates. Bars are mean +/- standard deviation of the mean. **(F)** TSS
1238 usage distribution based on insert length of TSS-seq reads generated from transformed
1239 libraries. Dots represent three biological replicates. Bars are mean +/- standard
1240 deviation of mean. Distributions are similar to the distributions in **(E)**. Note that primer
1241 extension will blur usage into adjacent upstream position due to some level of non-
1242 templated addition of C to RNA 5’ ends. **(G)** Heat scatter plots of Coefficient of Variation
1243 (CV, y axis) versus total RNA sequencing reads per promoter variant in each of three
1244 Pol II MASTER libraries. A cutoff of CV=0.5 was used to filter out variants with higher
1245 variance.

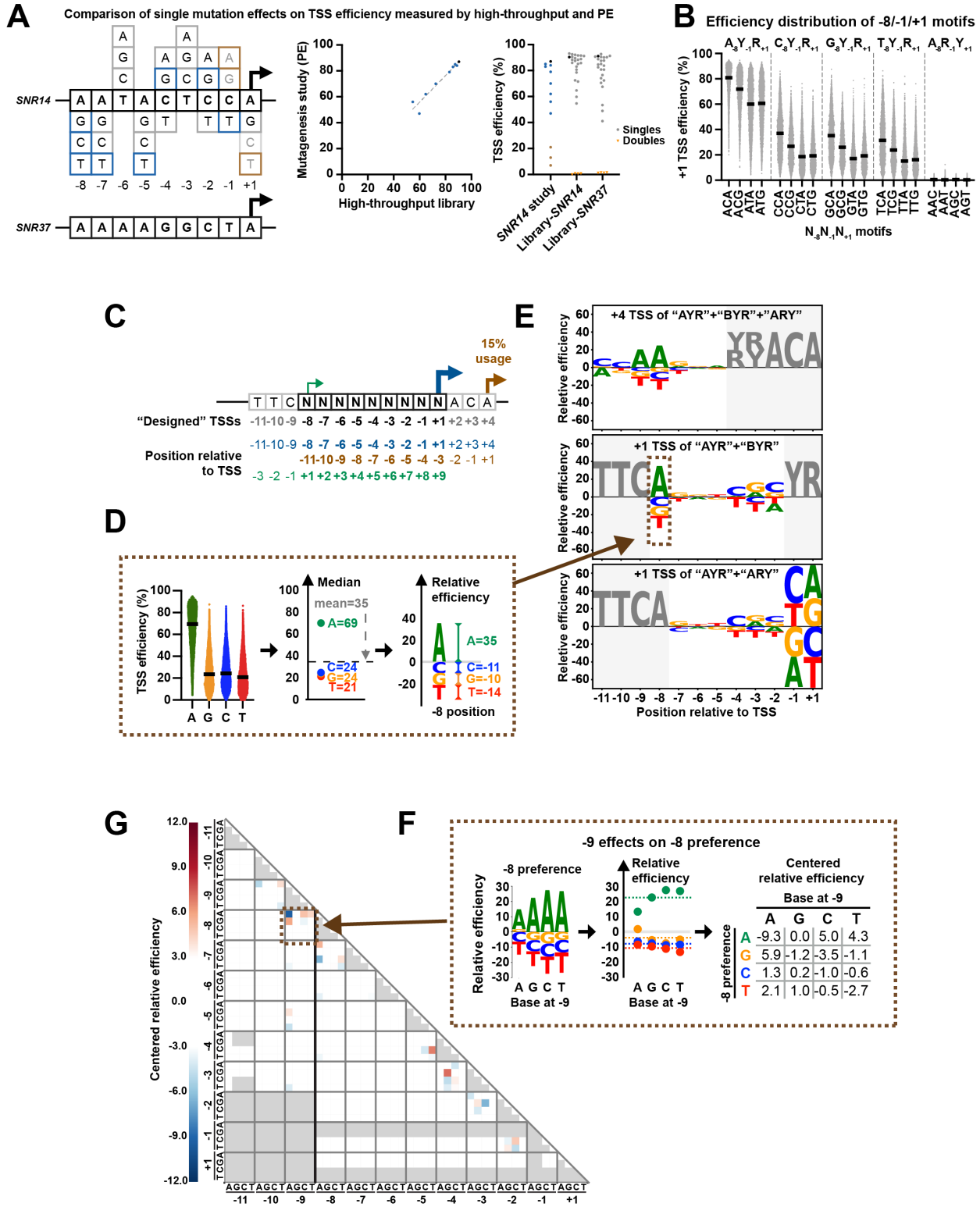


Figure 2. TSS and surrounding sequences modulate initiation efficiency over a wide range. (A) Comparison of single base mutation effects on TSS efficiency measured by Pol II MASTER and primer extension. (Left) Sequences of *SNR14* and

1250 *SNR37* TSS regions (in black boxes, including positions between -8 to +1 relative to
1251 TSS) and all possible single substitutions of *SNR14* TSS region. Single substitutions
1252 included in both a prior *SNR14* mutagenesis study [22] and Pol II MASTER libraries are in
1253 blue while those lacking in our study are in brown. Additional substitutions analyzed
1254 here are in gray. (Middle) High correlation of TSS efficiency measured by Pol II
1255 MASTER and primer extension. Mutation classes color coded as on left. (Right) Range
1256 of effects of single base substitutions on TSS efficiency for *SNR14*- and *SNR37*-related
1257 sequences. Mutation classes color coded as on left. Single substitutions absent from
1258 Pol II MASTER because of library design (R_{-1} and Y_{+1}) were expected to have super low
1259 TSS efficiencies. Double substitutions of *SNR14* and *SNR37* TSS region included in Pol
1260 II MASTER "ARY" library are shown as orange inverted triangles and show almost no
1261 efficiency. **(B)** Pol II initiation shows a strong preference for A_{-8} and $C_{-1}A_{+1}$ containing
1262 variants. All promoter variants were divided into 20 groups defined by bases at positions
1263 -8, -1 and +1 relative to the designed +1 TSS, and their +1 TSS efficiencies were plotted
1264 as spots. Lines represent median efficiencies of each group. **(C)** Definition of "designed"
1265 +1 TSS (designated as +1) and positions relative to this TSS (blue TSS arrow and
1266 sequence). TSS usage generated from upstream or downstream of "designed" +1 TSS
1267 (green and brown TSS arrows and sequences, respectively) allows us to study
1268 sequence preferences at positions -11 to -9 and +2 to +9 relative to site of initiation. **(D)**
1269 Schematic illustrating how "relative efficiency" is calculated and visualized as a
1270 sequence logo in **E**. At a particular position relative to a particular TSS, first, all variants
1271 were divided into the four base subgroups defined by base at this position. Next,
1272 median values of each base group were extracted and centered based on the mean of
1273 all median values. The centered median values were defined as "relative efficiencies",
1274 representing preferences for bases at this particular position. Finally, relative
1275 efficiencies of bases were visualized as sequence logos. Positive and negative values
1276 indicate relatively preferred or less preferred bases, respectively. **(E)** Pol II initiation
1277 shows distinct sequence preference at positions around the TSS. Top panel was
1278 generated using datasets of designed +4 TSS deriving from "AYR", "BYR" and "ARY"
1279 libraries. Middle panel was generated using datasets of designed +1 TSS deriving from
1280 "AYR" and "BYR" libraries. Bottom panel was generated using datasets of designed +1
1281 TSS deriving from "AYR" and "ARY" libraries. Positions that contain fixed or not
1282 completely randomized bases are shown in grey. **(F)** Schematic illustrating how
1283 sequence interaction between positions is calculated and visualized as a heat map in **G**.
1284 Using -9/-8 positions as an example, relative efficiencies at position -8 were calculated
1285 when different bases were present at position -9. Next, relative efficiencies of each base
1286 were centered based on the mean of all relative efficiencies of a particular base. After
1287 centralization, negative and positive values indicate negative and positive interactions.
1288 Interaction scores for two sequence positions are read at the intersection of the *x* and *y*
1289 axes labeled by base and position. That means, for example, -9.3 at -9A/-8A and 5.9 at
1290 -9A/-8G intersections indicate having a -9A decreases A preference and increases G
1291 preference at position -8, respectively. Finally, the centered relative efficiencies matrix
1292 was visualized as a heat map to represent the interaction between examined positions.
1293 **(G)** Sequence interactions are mainly observed at neighboring positions. Red and blue
1294 indicate positive and negative interactions, respectively. Missing values are shown in
1295 grey. Interactions related to positions -11 to -9 were calculated using datasets of

1296 designed +4 TSS deriving from “AYR”, “BYR” and “ARY” libraries. Other interactions
1297 were calculated using datasets of designed +1 TSS deriving from “AYR” and “BYR”
1298 libraries.

1299

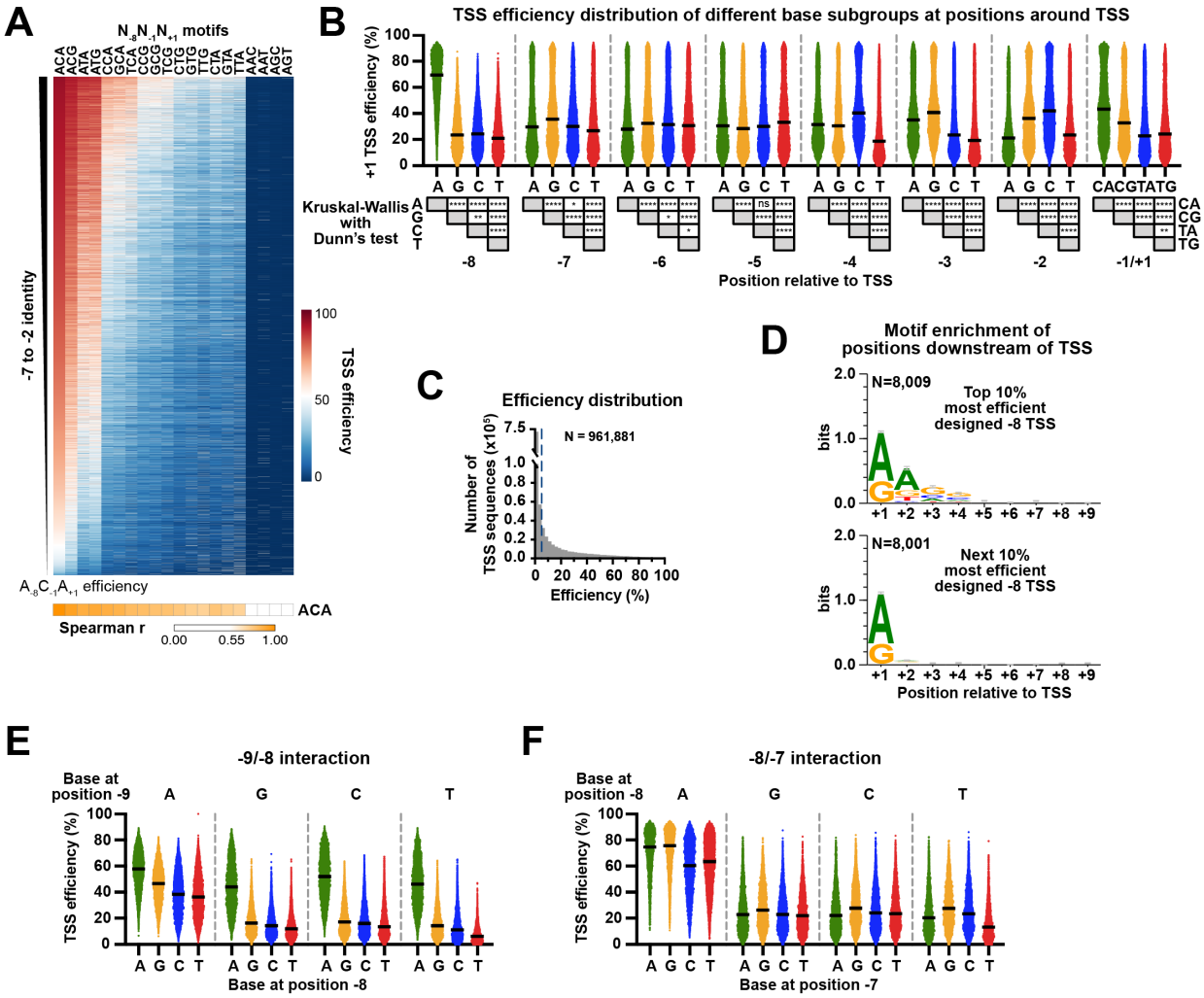


Figure S2. Surrounding sequence of TSSs modulates initiation efficiency. (A) +1 TSS efficiency of all -7 to -2 sequences within each $N_{-8}N_{-1}N_{+1}$ motif in WT, rank ordered by efficiency of their $A_{-8}C_{-1}A_{+1}$ version, is shown as a heat map. The x-axis is ordered based on median efficiency 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. **(B)** 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. ****, $P \leq 0.0001$; ***, $P \leq 0.001$; **, $P \leq 0.01$; *, $P \leq 0.05$. **(C)** 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. **(D)** $A_{+2}G_{+3}G_{+4}$ 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. Sequence logos were

1319 generated using WebLogo 3. Bars represent an approximate Bayesian 95% confidence
1320 interval. **(E)** An A at position -9 results in different sequence preferences at position -8.
1321 The dataset of designed +4 TSSs deriving from “AYR”, “BYR” and “ARY” libraries was
1322 used to detect the -9/-8 interaction. All variants were divided into 16 subgroups defined
1323 by bases at positions -9 and -8 relative to designed +4 TSS, and then their TSS
1324 efficiencies were plotted. Lines represent median values of subgroups. **(F)** An A at
1325 position -8 results in different sequence preferences at position -7. The dataset of
1326 designed +1 TSSs deriving from “AYR” and “BYR” libraries was used to detect -8/-7
1327 interaction. Calculations same as -9/-8 interaction described in **E**.

1328

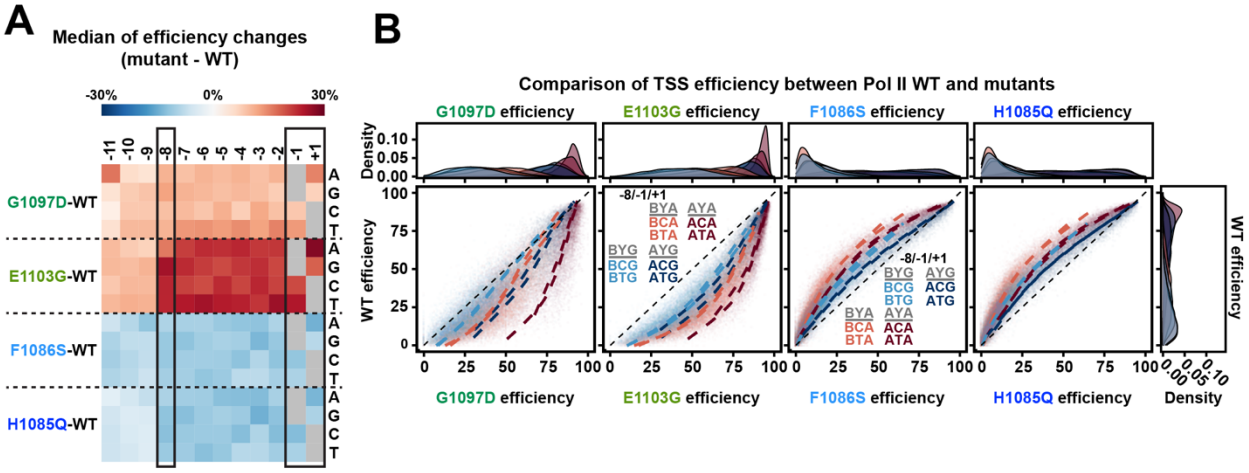
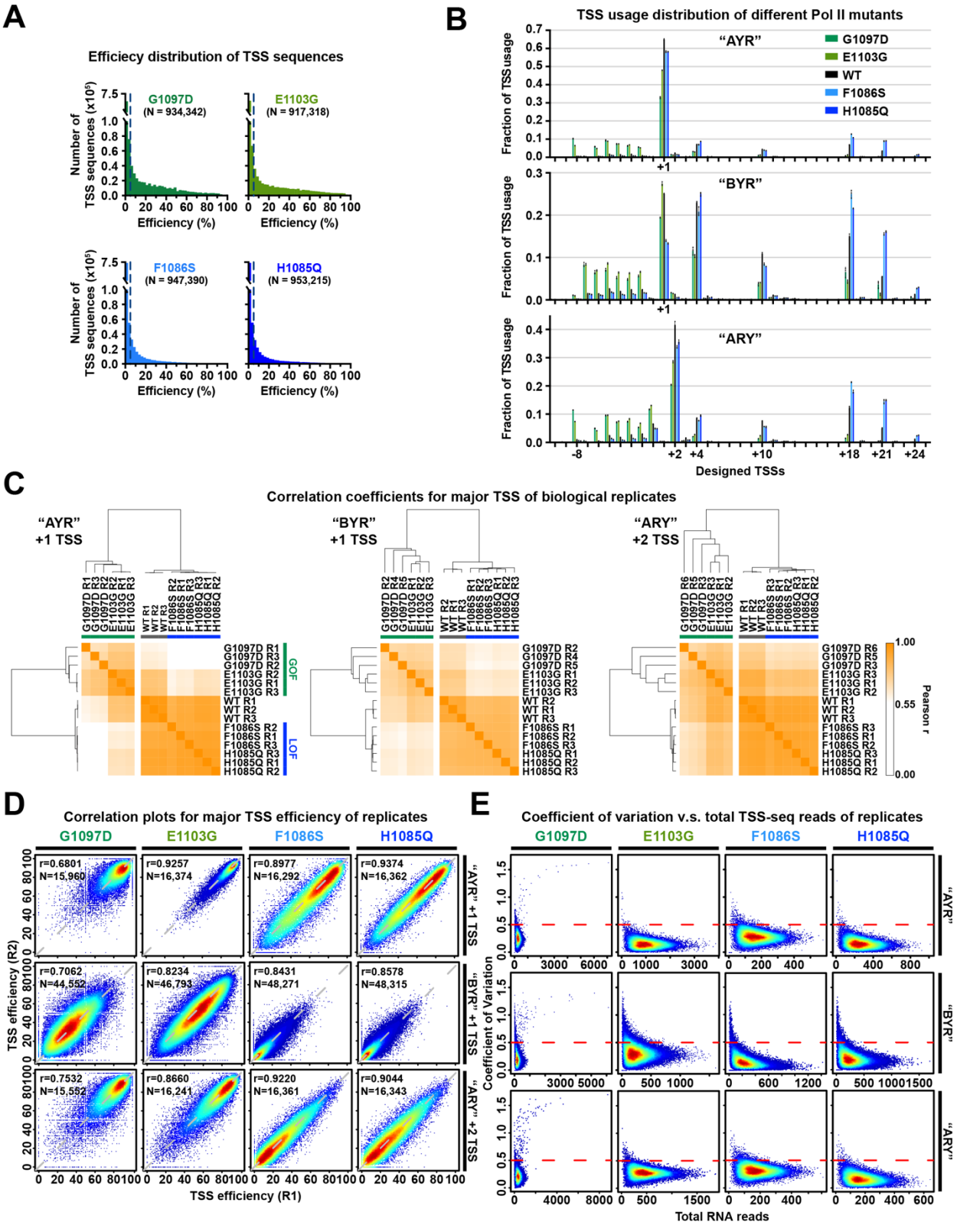


Figure 3. Pol II mutants alter TSS efficiency for all possible TSS motifs while showing selective effects for base at +1. (A) Pol II mutants alter TSS efficiencies across all motifs, corresponding to their direction of change to Pol II catalytic activity *in vitro*. TSS efficiency changes for each TSS 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. **(B)** 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 of 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
Efficiecy distribution of TSS sequences
G1097D
(N = 934,342)
E1103G
(N = 917,318)
F1086S
(N = 947,390)
H1085Q
(N = 953,215)
Number of TSS sequences (x10^5)
Efficiency (%)
B
TSS usage distribution of different Pol II mutants
"AYR"
"BYR"
"ARY"
G1097D
E1103G
WT
F1086S
H1085Q
Fraction of TSS usage
+1
-8
+2
+4
+10
+18
+21
+24
Designed TSSs
C
Correlation coefficients for major TSS of biological replicates
"AYR"
+1 TSS
"BYR"
+1 TSS
"ARY"
+2 TSS
GOF
LOF
Pearson r
D
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)
E
Coefficient of variation v.s. total TSS-seq reads of replicates
Coefficient of Variation
Total RNA reads

1352

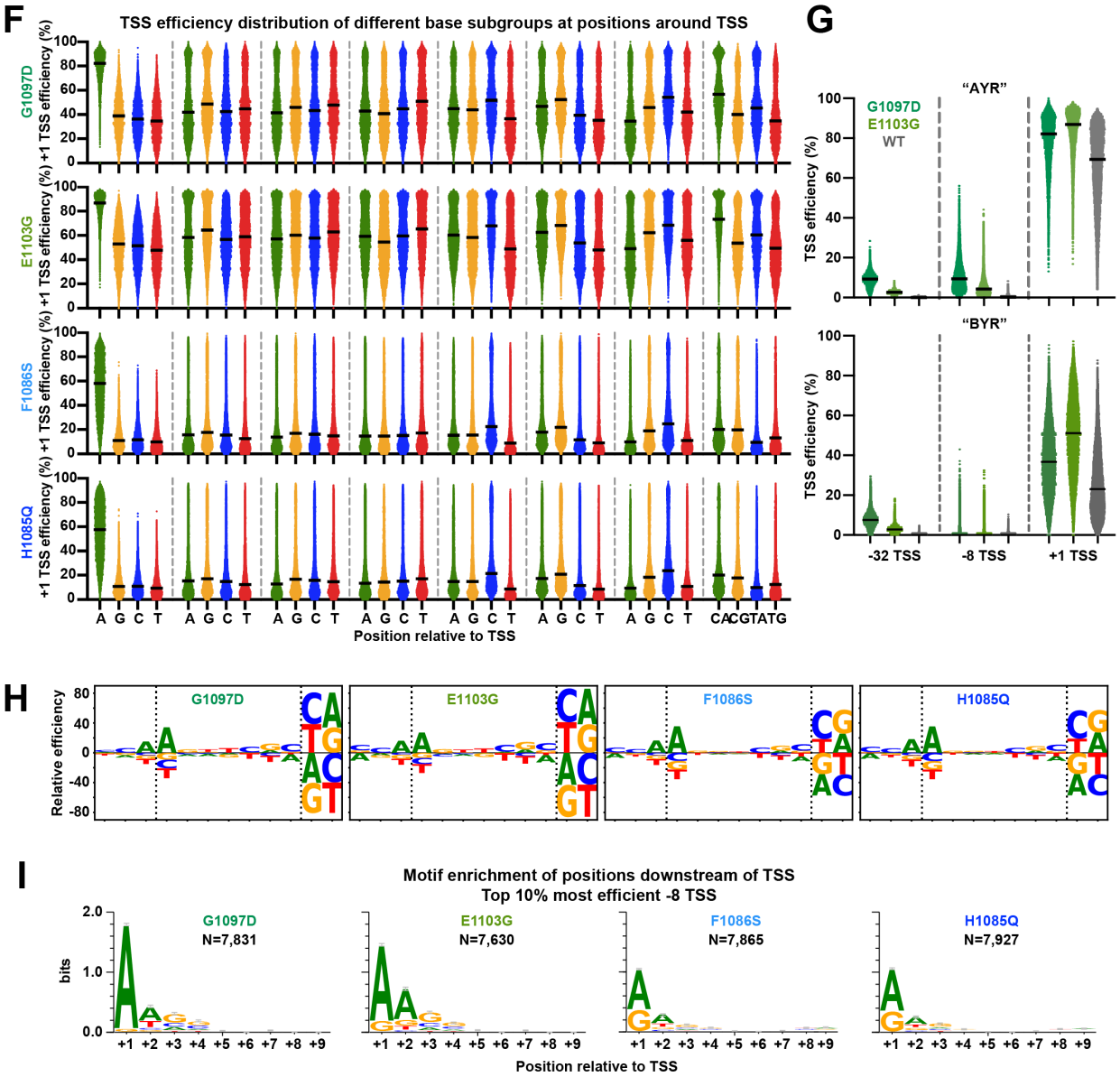


Figure S3. (A) Histograms showing the distribution of measured efficiencies for all designed -8 to +4 TSSs of all promoter variants deriving from "AYR", "BYR" and "ARY" libraries in Pol II mutants. Dashed lines mark the 5% efficiency cutoff used to designate a TSS as active. Total numbers of TSS sequences are shown. **(B)** 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. **(C)** Hierarchical clustering of Pearson correlation coefficients of TSS efficiencies for the 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. **(D)** 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. **(E)** Plots

1367 of the coefficient of variation (CV) versus the total RNA reads for three yeast replicates
1368 in Pol II mutants. The red dashed lines mark the CV=0.5 cutoff, which was chosen as an
1369 arbitrary cutoff for variants showing reasonable reproducibility across three biological
1370 replicates. G1097D replicates contain outliers because these slow growing strains are
1371 susceptible to genetic suppressors. These outliers are filtered out using our CV cutoff
1372 because of high CVs caused by suppressor only existing in one of three biological
1373 replicates. **(F)** TSS efficiency distributions of designed +1 TSSs of Pol II mutants for
1374 base subgroups at individual positions relative to +1. Identical analysis as in **Figure**
1375 **S2B** for WT was performed for Pol II mutant libraries. **(G)** Pol II GOF G1097D showed
1376 greater increase in efficiency than GOF allele E1103G at upstream TSSs (designed -32
1377 and -8 TSSs), while E1103G showed stronger effects at designed +1 TSS than
1378 G1097D. This may indicate that where Pol II is in the scanning process may affect
1379 efficiency. **(H)** Pol II initiation sequence preference in Pol II mutants. Identical analysis
1380 as in **Figure 2E** for WT was performed for Pol II mutant libraries. Sequence logos were
1381 generated using multiple datasets and the dashed lines indicate divisions in datasets
1382 used to generate them. Specifically, preferences at positions -11 to -9 were generated
1383 using datasets of designed +4 TSS deriving from “AYR”, “BYR” and “ARY” libraries.
1384 Preferences at positions -8 to -2 were generated using datasets of designed +1 TSS
1385 deriving from “AYR” and “BYR” libraries. Preferences at positions -1 and +1 were
1386 generated using datasets of designed +1 TSS deriving from “AYR” and “ARY” libraries.
1387 **(I)** Motif enrichment for top the 10% most efficient -8 TSSs for Pol II mutants. Identical
1388 motif enrichment analysis as in **Figure S2D** top panel for WT was performed to Pol II
1389 mutant libraries. Numbers (N) of variants assessed are indicated. Bars represent an
1390 approximate Bayesian 95% confidence interval.

1391

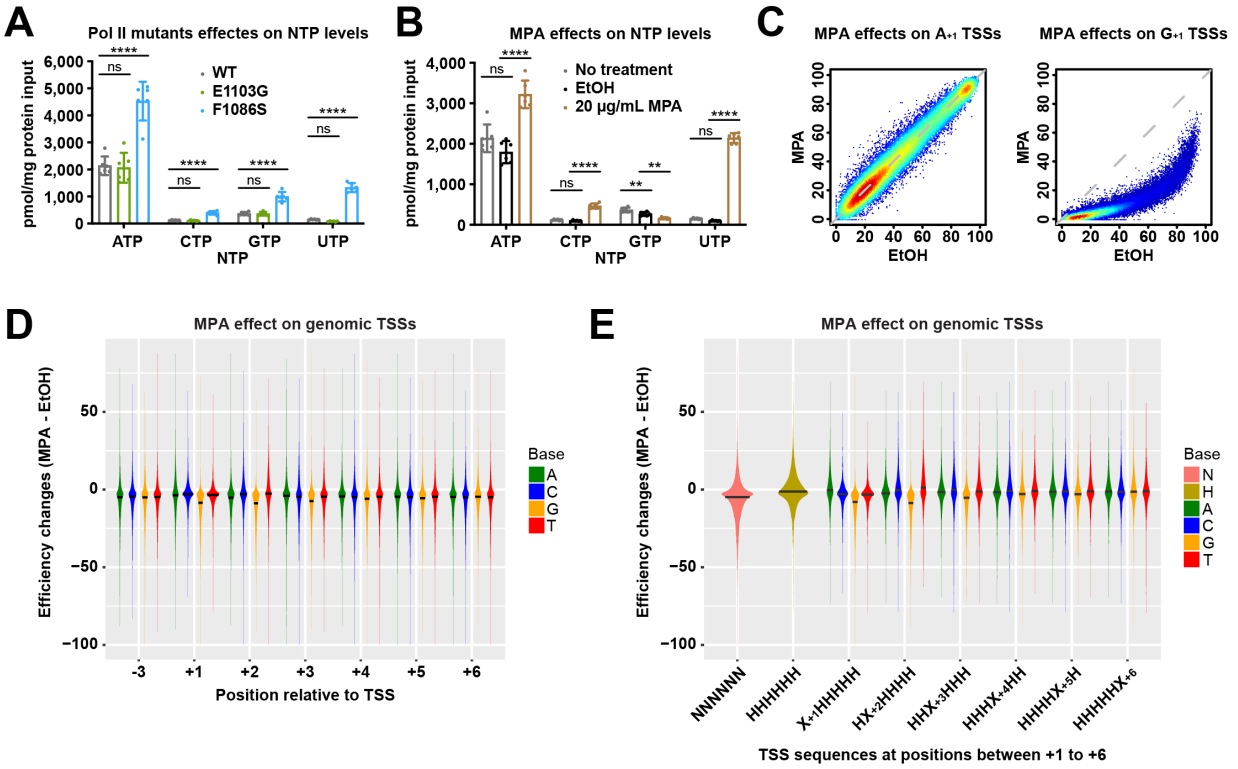
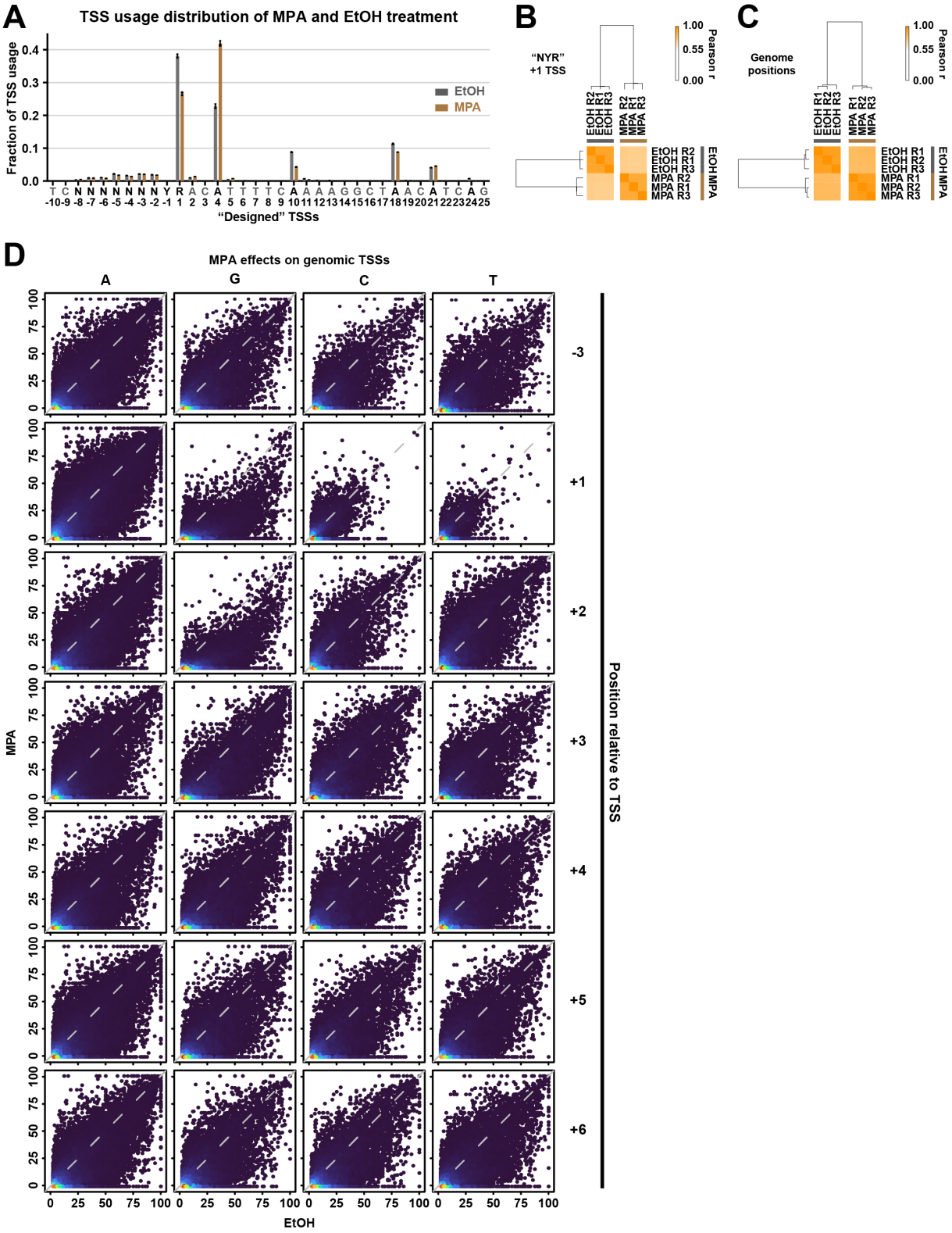


Figure 4. Pol II initiation is sensitive to NTP pools. (A) NTP levels measured in WT, Pol II E1103G (GOF) and Pol II F1086S (LOF) mutants. Statistical analyses by Ordinary one-way ANOVA with Dunnett's multiple comparisons test for each NTP leave are shown. ****, P ≤ 0.0001; ***, P ≤ 0.001; **, P ≤ 0.01; *, P ≤ 0.05. **(B)** NTP levels measured in WT, WT treated with 100% ethanol (the solvent for MPA) and WT treated with 20 ug/ml MPA. Statistical analyses by RM one-way ANOVA with Dunnett's multiple comparisons test for each NTP leave are shown. ****, P ≤ 0.0001; ***, P ≤ 0.001; **, P ≤ 0.01; *, P ≤ 0.05. **(C)** MPA selectively decreases TSS efficiencies of +1G TSSs (right) but shows no effects on +1A TSSs (left) in Pol II MASTER libraries. **(D)** MPA alters TSS efficiencies at genomic TSS positions. TSS efficiency change shown as difference between efficiency in EtOH and MPA treatment for all TSSs ≥ 2% efficiency in EtOH conditions within the 25%-75% of TSS distributions (see Methods). Bars on violin plots indicate medians computed based on density estimates. (**E**) Comparison of TSSs of any composition for first six nucleotides (NNNNNN) with those that lack G (HHHHHH, H=not G) or subsets of TSSs with A,C,G,T at positions +1 to +6. Bars on violin plots are as in (D). All C,G,T datasets for all positions 1-5 are distinct from A as a baseline comparison (Wilcoxon rank sum test with continuity correction). Position 6 each base is not significantly different from A.



1411

1412 **Figure S4. High of reproducibility of TSS usage and efficiency upon MPA**
1413 **treatment. (A)** TSS usage distributions at designed -10 to +25 TSSs in WT "NYR"
1414 library (mixed AYR and BYR libraries) treated with 100% ethanol or with 20 ug/ml MPA.

1415 MPA treatment shifted TSS usage downstream relative to EtOH treatment. Dots
1416 represent three biological replicates. Bars are mean +/- standard deviation of the mean.
1417 **(B)** Hierarchical clustering of Pearson correlation coefficients of TSS efficiencies for
1418 designed +1 TSS for three biological replicates for MPA or EtOH treatment, illustrated
1419 as a heat map. **(C)** Hierarchical clustering of Pearson correlation coefficients of TSS
1420 efficiencies for all genome positions within defined promoter windows with >=3 reads in
1421 each replicate, illustrated as a heat map. **(D)** Correlation plots for combined biological
1422 replicates for TSS efficiency upon MPA treatment (*y* axes) versus EtOH treatment (*x*
1423 axes) for all TSSs ≥2% efficiency in the 25%-75% of the distribution for a curated set of
1424 5979 yeast promoters (see Methods). TSSs are separated into groups depending on
1425 base identity at positions -3 (control) or positions +1 to +6.

1426

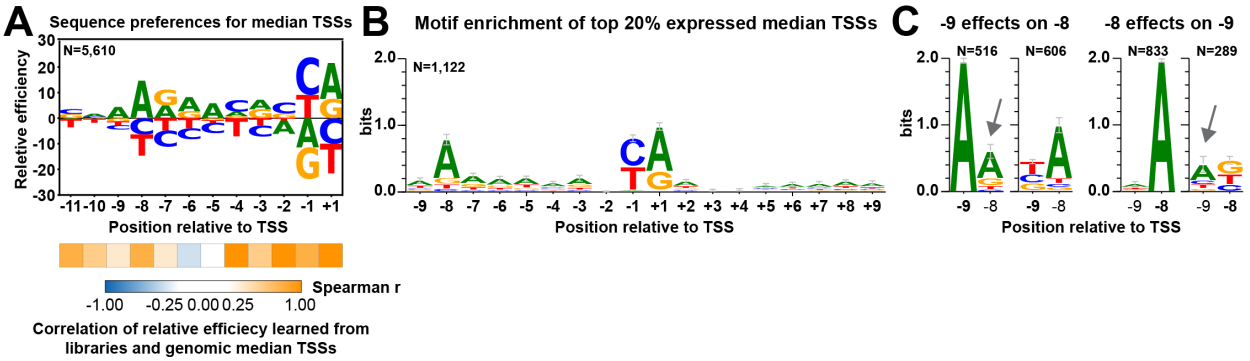


Figure 5. Learned initiation preferences are predictive of TSS efficiencies at genomic promoters. (A) Sequence preferences determined from TSSs representing the median of the distribution of usage from a set of 5979 yeast promoters ("median TSSs") are congruent with library determined TSS preferences, except there is preference for A at positions -7 to -5 learned from genomic data. Sequence context and TSS efficiency of median TSSs were extracted from genomic TSS-seq data (GSE182792) [27]. 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. Bars represent an approximate Bayesian 95% confidence interval. **(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. Bars represent an approximate Bayesian 95% confidence interval.

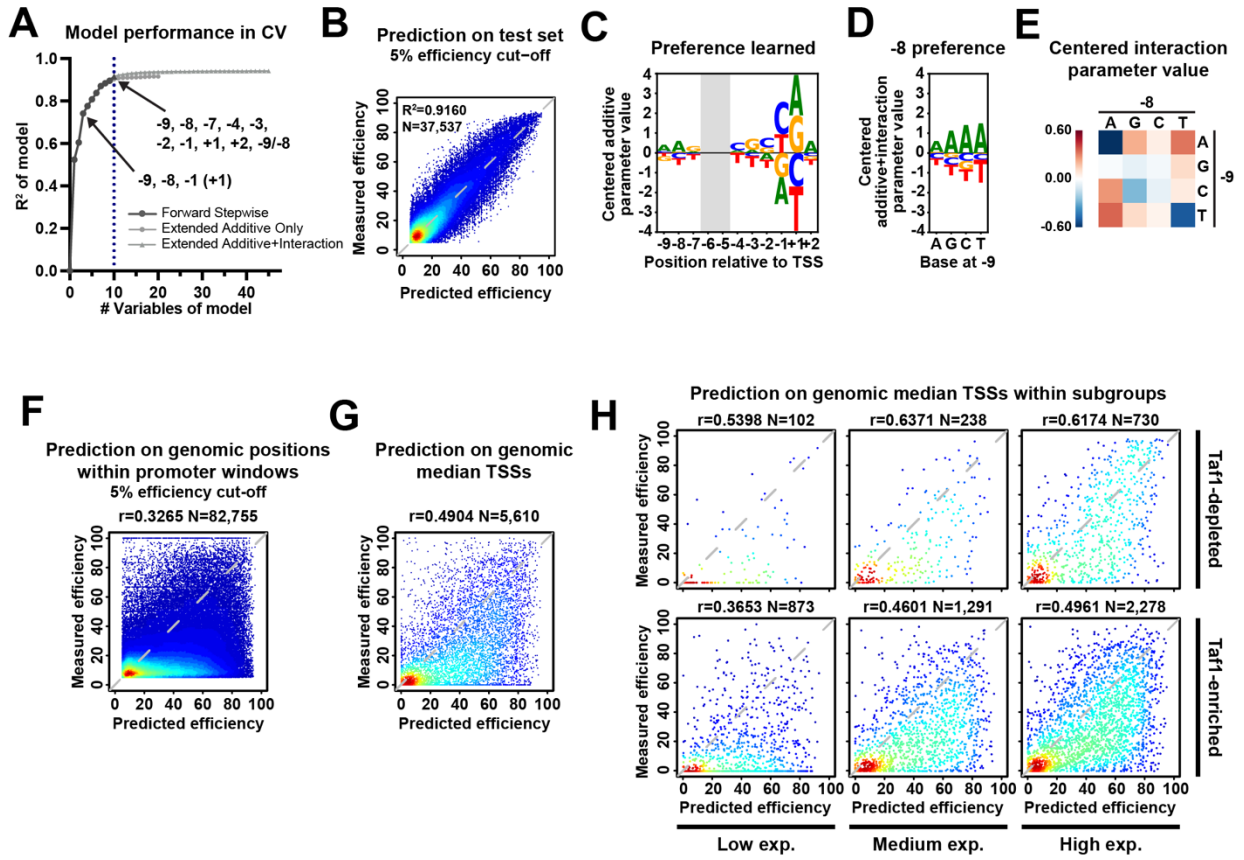


Figure 6. Logistic regression model of DNA sequence contribution to TSS efficiency. (A) Regression modeling identifies key DNA sequences and interactions contributing to TSS efficiency. Nine additive parameters plus one interaction were selected for the final model. Dots represents average R^2 obtained 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. **(B** to **E)** 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. (**B**) 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. (**C**) A sequence logo of centered additive parameters. The coefficients for bases at a particular position was centered and visualized as a sequence logo. (**D**) 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. (**E**) A heat map of centered parameters for -9/-8 interaction illustrating how bases at one position affect preference at another position. **(F** to **G)** Efficiency prediction for positions within known promoter windows in WT shows overall over-prediction. Scatterplots of comparison of measured and predicted TSS

1469 efficiencies of all positions (with a 5% efficiency cut-off) (**F**) or median TSSs (**G**) within
1470 5979 known genomic promoter windows [21] with available measured efficiency. **(H)**
1471 Model shows better performance on Taf1-depleted promoters and promoters with
1472 medium to high expression. Scatterplots of comparison of measured and predicted TSS
1473 efficiency of median TSSs subgrouped by promoter classes and expression levels.
1474 Expression levels of genomic promoters are defined based on their total TSS-seq reads
1475 in the promoter window in the examined datasets: low, [0, 200); medium, [200, 1000);
1476 high, [1000, max). Pearson r and number (N) of compared variants are shown.

1477

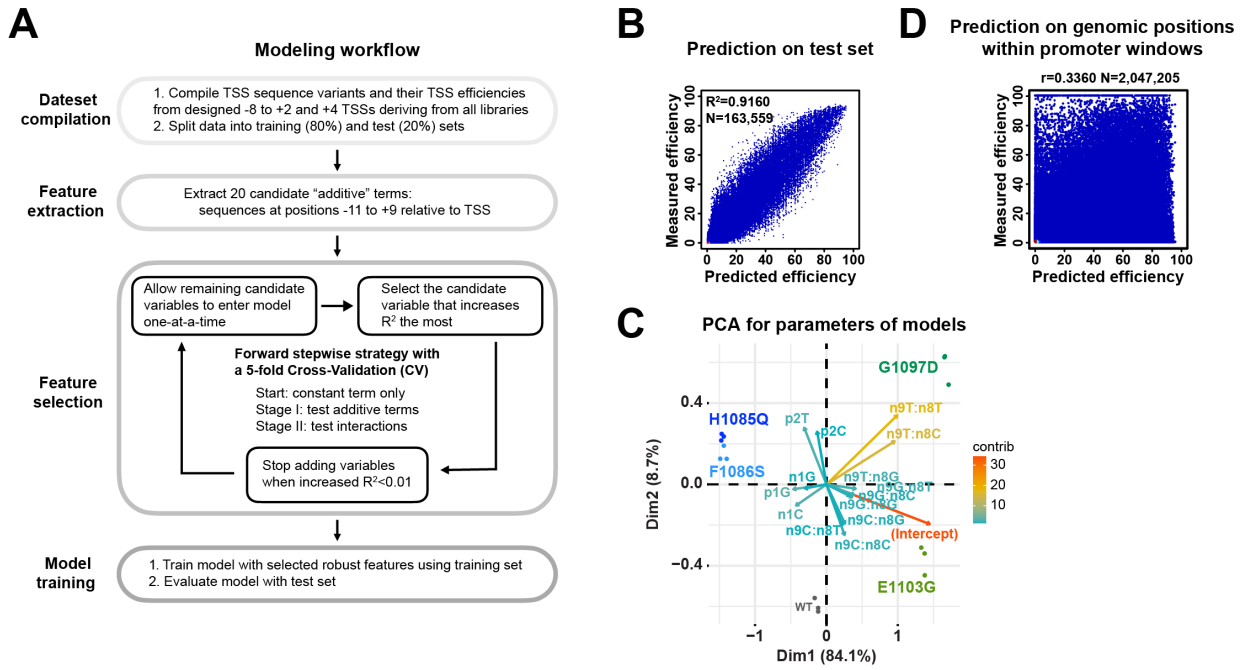


Figure S5. (A) Overview of the TSS efficiency modeling process. (1) Variants including designed -8 to +2 and +4 TSSs deriving from "AYR", "BYR" and "ARY" libraries with available TSS efficiency were pooled for modeling. (2) Sequences at positions -11 to +9 relative to TSS of each variant were extracted. (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)** 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. **(C)** PCA analysis for parameters of models trained by the individual replicates of WT and Pol II mutant. Parameters for replicates cluster more closely than parameters for different Pol II mutants, indicating that modeling captured features of different Pol II groups and models are not overfit. 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 S3I**, where G1097D and E1103G differentially altered +2 sequence enrichment. **(D)** A scatterplot of comparison of measured and predicted TSS efficiencies of all positions within 5979 known genomic promoter windows [21] 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.

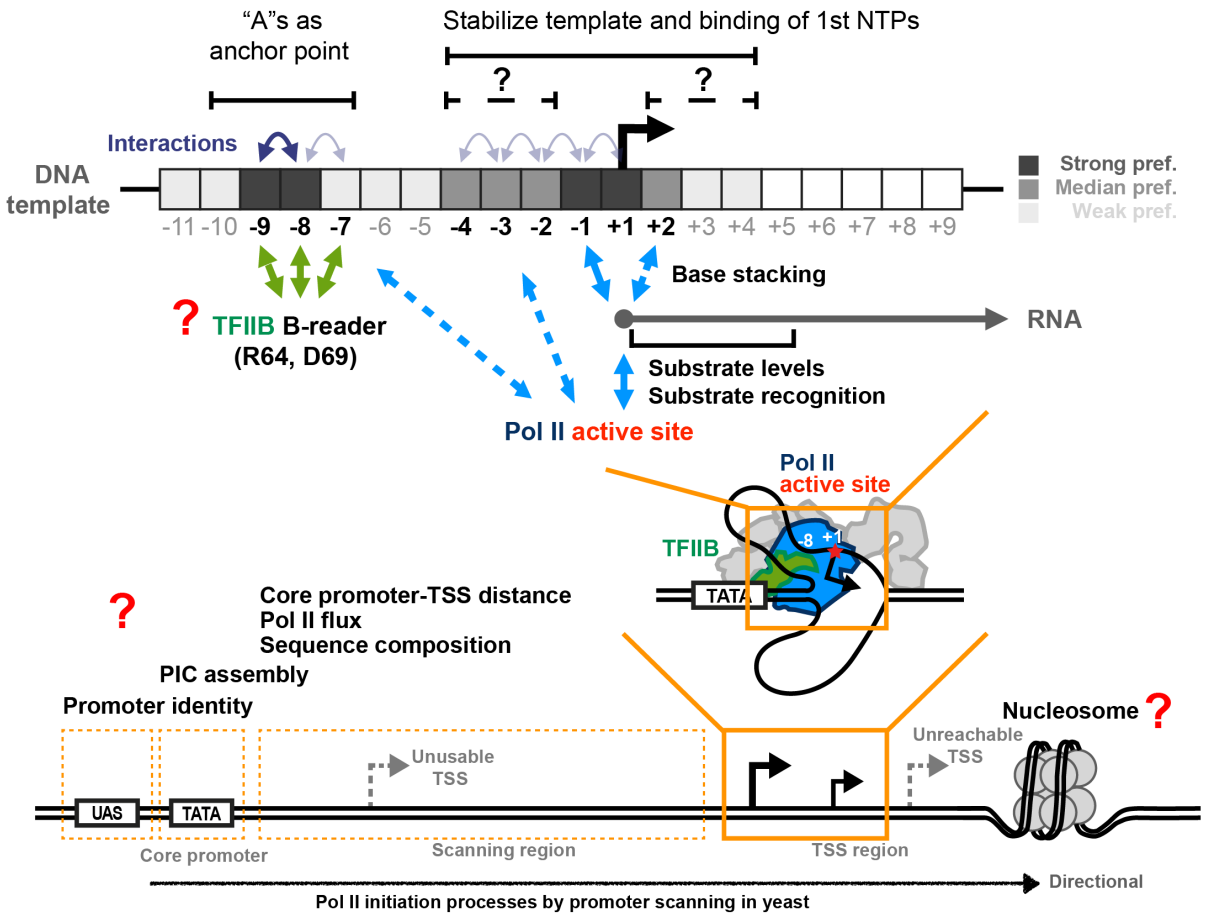


Figure 7. Model for TSS sequence preference regulated by multiple mechanisms. Top panels show determined contribution of sequence at positions around TSS and proposed mechanisms. Two major groups of positions around TSS contribute to TSS selection: bases around TSS (actual initiating site) and bases 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 may 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, as proposed [47]. These preferences are reflected 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.