Promoter-sequence determinants and structural basis of primer-dependent transcription initiation in *Escherichia coli*

Kyle S. Skalenko[1,2], Lingting Li[3], Yuanchao Zhang[4], Irina O. Vvedenskaya[1,2], Jared T. Winkelman[1,2,5], Alexander Cope[1], Deanne M. Taylor[4], Premal Shah[1], Richard H. Ebright[2,5], Justin B. Kinney[6], Yu Zhang[3], and Bryce E. Nickels[1,2,*]

[1] Department of Genetics, Rutgers University, Piscataway, NJ 08854, USA.
[2] Waksman Institute, Rutgers University, Piscataway, NJ 08854, USA.
[3] Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Xuhui, Shanghai, China, 200032
[4] Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA 19041, USA.
[5] Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, NJ 08854, USA.
[6] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA.

* Corresponding author, Bryce E. Nickels
Email: bnickels@waksman.rutgers.edu

Abstract

Chemical modifications of RNA 5′ ends enable "epitranscriptomic" regulation, influencing multiple aspects of RNA fate. In transcription initiation, a large inventory of substrates compete with nucleoside triphosphates (NTPs) for use as initiating entities, providing an *ab initio* mechanism for altering the RNA 5′ end. In *Escherichia coli* cells, RNAs with a 5′-end hydroxyl are generated by use of dinucleotide RNAs as primers for transcription initiation, "primer-dependent initiation." Here we use massively systematic transcript end readout ("MASTER") to detect and quantify RNA 5′ ends generated by primer-dependent initiation for ~4^{10} (~1,000,000) promoter sequences in *E. coli*. The results show primer-dependent initiation in *E. coli* involves any of the 16 possible dinucleotide primers and depends on promoter sequences in, upstream, and downstream of the primer binding site. The results yield a consensus sequence for primer-dependent initiation, $Y_{TSS-2}N_{TSS-1}N_{TSS}W_{TSS+1}$, where TSS is the transcription start site, $N_{TSS-1}N_{TSS}$ is the primer binding site, Y is pyrimidine, and W is A or T. Biochemical and structure-determination studies show that the base pair (nontemplate-strand base:template-strand base) immediately upstream of the primer binding site ($Y:R_{TSS-2}$, where R is purine) exerts its effect through the base on the DNA template strand (R_{TSS-2}) through inter-chain base stacking with the RNA primer. Results from analysis of a large set of natural, chromosomally-encoded *E. coli* promoters support the conclusions from MASTER. Our findings provide a mechanistic and structural description of how TSS-region sequence hard-codes not only the TSS position, but also the potential for epitranscriptomic regulation through primer-dependent transcription initiation.

Introduction

In transcription initiation, the RNA polymerase (RNAP) holoenzyme binds promoter DNA by making sequence-specific interactions with core promoter elements and unwinds a turn of promoter DNA forming an RNAP-promoter open complex (RPo) containing a single-stranded "transcription bubble." Next, RNAP selects a transcription start site (TSS) by placing the start-site nucleotide and the next nucleotide of the "template DNA strand" into the RNAP active-center product site ("P site") and addition site ("A site"), respectively, and binding an initiating entity in the RNAP active-center P site (Figure 1A). RNAPs can initiate transcription using either a primer-independent or primer-dependent mechanism (1-10). In primer-independent initiation, the initiating entity (typically a nucleoside triphosphate, NTP) base pairs to the template-strand nucleotide in the RNAP active-center P site (TSS; Figure 1A). In primer-dependent transcription initiation, the 3′ nucleotide of a 2-, 3-, or 4-nucleotide RNA primer (di-, tri-, or tetranucleotide primer, respectively) base pairs to the template-strand nucleotide in the RNAP active-center P site, and the 5′ nucleotide of the primer base pairs to the template-strand nucleotide in the P-1, P-2, or P-3 site (TSS-1, TSS-2, and TSS-3, respectively; Figure 1A).

In *Escherichia coli* cells, primer-dependent transcription initiation occurs during stationary-phase growth and modulates the expression of genes involved in biofilm formation (9-11). RNAs generated by primer-dependent initiation in *E. coli* contain a 5′-end hydroxyl (5′-OH), indicating that the primers incorporated at the RNA 5′ end also contain a 5′-OH (9). Available data suggests that most primer-dependent initiation in *E. coli* involves use of dinucleotide primers, most frequently UpA and GpG (9-11). However, direct evidence that dinucleotides serve as the predominant initiating entity in primer-dependent initiation has not been presented. In addition, apart from the sequences complementary to the primer, "the primer binding site," promoter-sequence determinants for primer-dependent initiation have not been defined.

Here we adapt a massively parallel reporter assay to monitor primer-dependent initiation in *E. coli*. The results provide a complete inventory of the RNA 5′-end sequences generated by primer-dependent initiation in *E. coli* and define the critical promoter-sequence determinants for primer-dependent initiation. The results demonstrate that most, if not all, primer-dependent initiation in *E. coli* involves use of a dinucleotide as the initiating entity and identify a consensus sequence for primer-dependent initiation, $Y_{TSS-2}N_{TSS-1}N_{TSS}W_{TSS+1}$, where TSS is the transcription start site, $N_{TSS-1}N_{TSS}$ is the primer binding site, Y is pyrimidine, and W is A or T. We further demonstrate that sequence information at the position immediately upstream of the primer binding site resides exclusively in the template strand of the transcription bubble (R_{TSS-2}, where R is purine). We report crystal structures of transcription initiation complexes containing dinucleotide primers that reveal the structural basis for a

purine at the template-strand position immediately upstream of the primer binding site (R_{TSS-2}): namely, more extensive, and likely more energetically favorable, base stacking between the template-strand base and the primer 5′ base.

Results

Use of massively systematic transcript end readout, "MASTER," to monitor primer-dependent initiation in E. coli

To define, comprehensively, the promoter-sequence determinants for primer-dependent initiation in *E. coli*, we modified a massively parallel reporter assay previously developed in our lab termed massively systematic transcript end readout, "MASTER" (12, 13), in order to detect both primer-independent and primer-dependent initiation, to differentiate between primer-independent and primer-dependent initiation, and to define primer lengths in primer-dependent initiation (Figure 1B).

MASTER involves construction of a promoter library that contains up to 4^{11} (~4,000,000) barcoded sequences, production of RNA transcripts from the promoter library, and analysis of RNA barcodes and RNA 5′ ends using high-throughput sequencing (5′ RNA-seq) to define, for each RNA product, the template that produced the RNA and the sequence of the RNA 5′ end (Figure 1B) (12-14). The 5′ RNA-seq procedure used in MASTER relies on ligation of single-stranded oligonucleotide adaptors to RNAs containing a 5′-end monophosphate (5′-p) (13). In previous work, we treated RNAs with RNA 5′ Pyrophosphohydrolase (Rpp), which converts a 5′-end triphosphate (5′-ppp) to a 5′-p; this procedure specifically enables detection of the 5′-ppp bearing RNAs generated by primer-independent initiation (12, 14-16). Here, we treated RNAs, in parallel, with Rpp to detect RNAs generated by primer-independent initiation and with Polynucleotide Kinase (PNK), which converts a 5′-OH to a 5′-p, to detect RNAs generated by primer-dependent initiation (Figure 1B). By comparing the results from Rpp and PNK reactions we quantify, for each promoter sequence in the library, the relative efficiencies of primer-independent and primer-dependent initiation, and primer lengths for primer-dependent initiation.

In the present work, we used a MASTER template library containing 4^{10} (~1,000,000) sequence variants at the positions 1-10 base pairs (bp) downstream of the -10 element of a consensus σ^{70}-dependent promoter (p*lac*CONS-N10; Figure 1B). The randomized segment of p*lac*CONS-N10 contains the full range of TSS positions for *E. coli* RNAP observed in previous work (i.e., TSS positions located 6, 7, 8, 9, and 10 bp downstream of the promoter -10 element; Figure 2A). We introduced the p*lac*CONS-N10 library into *E. coli*, grew cells to stationary phase (the phase in which primer-dependent initiation has been observed in previous work; 9), isolated total cellular RNA, and analyzed RNAs generated from each

promoter sequence in the library by 5′ RNA-seq. The results provide complete inventories of RNA 5′ ends generated by primer-independent initiation and primer-dependent initiation in stationary-phase *E. coli* cells.

Primer-dependent initiation: 5′-end positions

Our results define distributions of 5′-end positions of the 5′-ppp RNAs generated by primer-independent initiation and the 5′-OH RNAs generated by primer-dependent initiation for transcription in stationary-phase *E. coli* cells (Figure 2B).

The distributions of 5′-end positions for primer-independent initiation show 5′-end positions (TSS positions) ranging from 6-10 bp downstream of the promoter -10 element, with a mean 5′-end position ~7.5 bp downstream of the promoter -10 element (Figure 2B, top). The range, the mean, and the distribution shape closely match those previously observed for primer-independent initiation for cells in exponential phase (12).

The distributions of 5′-end positions for primer-dependent initiation show 5′-end positions ranging from 5-9 bp downstream of the promoter -10 element, with a mean 5′-end position ~6.8 bp downstream of the promoter -10 element (Figure 2B, bottom). The range, the mean, and the distribution shape closely match those for primer-independent initiation, but with a ~1 bp upstream shift.

Primer-dependent initiation: primer lengths

Comparison of the 5′-end distributions for primer-independent initiation (Figure 2B, top) vs. primer-dependent initiation (Figure 2B, bottom) indicates that, across all promoter sequences in the library, the 5′-end positions of RNAs generated by primer-independent initiation (mean position 7.54 ± 0.01 bp downstream of -10 element) and RNAs generated by primer-dependent initiation (mean position 6.75 ± 0.05 bp downstream of -10 element) differ by ~1 bp (0.71 ± 0.06 bp). Following the logic of Figure 1, based on the observed difference of almost exactly 1 bp in mean 5′-end position for primer-independent initiation vs. primer-dependent initiation, we infer that primer length in primer-dependent initiation in stationary-phase *E. coli* cells is almost always 2 nt. Computational modeling, using the distributions in Figure 2B, indicates that no more than ~2.5% of the observed primer-dependent initiation could involve primer lengths greater than 2-nt (Figure S1A). Consistent with these inferences, comparison of distributions of RNA 5′-end positions for primer-independent initiation *in vitro* vs. primer-dependent initiation *in vitro* with the dinucleotide primer UpA shows essentially the same ~1 bp upstream shift in distribution range, mode, and mean (Figures 2C, S1B).

Primer-dependent initiation: primer sequences

We next measured yields of 5′-OH RNAs generated by primer-dependent initiation with each of the 16 possible dinucleotide primers (Figure 3A). The results show that primer-dependent initiation occurs with all 16 dinucleotide primers. Highest levels of primer-dependent initiation are observed with the dinucleotide primers UpA and GpG, which account for ~27% and ~17%, respectively, of 5′-OH RNAs generated across all promoters in the library (Figure 3A, left). The other 14 dinucleotide primers each account for ~1% to ~8% of 5′-OH RNAs generated across all promoters in the library. Qualitatively similar results are obtained analyzing RNA products from promoters where the primer binding site is at positions 5-6, 6-7, 7-8, 8-9, and 9-10 bp downstream of the promoter -10 element (Figure 3A, right). The demonstration that primer-dependent initiation occurs with all 16 dinucleotides *in vivo* is new to this work, as is the demonstration that primer-dependent initiation occurs at the full range of TSS positions observed for primer-independent initiation *in vivo* (i.e., TSS positions located 6, 7, 8, 9, and 10 bp downstream of the promoter -10 element). The observation that UpA and GpG are preferentially used as primers *in vivo* is consistent with results of prior work (9, 10).

Primer-dependent initiation: promoter-sequence dependence, primer binding site

Analysis of the results for primer-dependent initiation, separately considering RNA products with 5′ ends complementary to the template and RNA products with 5′ ends non-complementary to the template, shows that the overwhelming majority of primer-dependent initiation in stationary-phase *E. coli* cells occurs at primer binding sites that have perfect template-strand complementarity to the 5′ and 3′ nucleotides of the dinucleotide primer (93.3 ± 0.4%; Figure 3B, bottom). This is true, across the entire promoter library, for each of the 16 possible dinucleotide primer sequences (73.9 ± 0.2% to 98.1 ± 0.01% of primer binding sites with perfect complementarity; Figure 3B, bottom), and for each of the major primer binding-site positions (Figure S2). Most of the limited non-complementarity observed involves the 5′ nucleotide of the dinucleotide primers CpG, UpG, CpU and UpU (24.2 ± 0.4%, 21.1 ± 0.6%, 10.3 ± 0.6%, and 10.1 ± 0.9%, respectively; Figures 3B, S2). Consistent with these results, analysis of the same promoter library *in vitro*, assessing primer-dependent initiation with the dinucleotide primer UpA, shows that the overwhelming majority of primer-dependent initiation likewise occurs at primer binding sites that have perfect template-strand complementarity to the 5′ and 3′ nucleotides of the dinucleotide primer for each of the major primer binding-site positions (84.1 ± 0.6%; Figures 3B, bottom right, S3). *In vitro* transcription experiments using heteroduplex templates (templates having non-complementary transcription-bubble nontemplate-strand and template-strand sequences) and the dinucleotide primer UpA show that the strong preference for perfect template-strand complementarity to the 5′ and 3′ nucleotides of the dinucleotide primer reflects a requirement for Watson-Crick base pairing of template-strand

nucleotides at positions TSS-1 and TSS with the 5′ and 3′ nucleotides of the dinucleotide RNA primer, respectively (Figure S4).

We conclude that primer-dependent initiation with a dinucleotide primer almost always involves a primer binding site having perfect template-strand complementarity to, and therefore able to engage in Watson-Crick base pairing with, the dinucleotide primer. This result is not completely unexpected. However, this point has not been demonstrated previously *in vivo*, and prior work *in vitro*, with tetranucleotide primers (17), had indicated that perfect template-strand complementarity to the primer may not be necessary for primer-dependent initiation with longer primers.

Primer-dependent initiation: promoter-sequence dependence, sequences flanking the primer binding site

The observed yields of 5′-OH RNA products from primer-dependent initiation in stationary-phase *E. coli* cells strongly correlate with the promoter sequences flanking the primer binding site (Figures 4, S5). Levels of primer-dependent initiation depend on the identities of base pairs up to 7 bases upstream of the primer binding site and up to 3 bases downstream of the primer binding site. The base pair (nontemplate-strand base:template-strand base) at the position immediately upstream of the primer binding site, position TSS-2, makes the largest contribution to the sequence dependence of primer-dependent initiation. Levels of primer-dependent initiation are higher for promoters with a nontemplate-strand pyrimidine (C or T) and template-strand purine (A or G) at position TSS-2 (Figure 4, left). A strong preference for a Y:R base pair at position TSS-2 is observed at each of the major TSS positions (i.e., the positions 6, 7, and 8 bp downstream of the promoter -10 element; Figure 4, left), The results further show that the base pair at the position immediately downstream of the primer binding site, position TSS+1, makes the second largest contribution to the sequence dependence of primer-dependent initiation; levels of primer-dependent initiation are higher for promoters with a T:A or A:T base pair at position TSS+1 (Figure 4, left). A strong preference for T:A or A:T at position TSS+1 is observed when the TSS is 6 bp downstream of the promoter -10 element, and a weaker preference is observed when the TSS is 7 or 8 bp downstream of the promoter -10 element (Figure 4, left). Base pairs at positions 3, 4, 5, 6, and 7 bp upstream of the primer binding site (TSS-3, TSS-4, TSS-5, TSS-6, and TSS-7) and at positions 2 and 3 bp downstream of the primer binding site (TSS+2 and TSS+3), make small, but significant, contributions to levels of primer-dependent initiation (Figure 4, left). The results define a global consensus sequence for primer-dependent initiation: $Y_{TSS-2}N_{TSS-1}N_{TSS}W_{TSS+1}$ ($Y:R_{TSS-2}N:N_{TSS-1}N:N_{TSS}W:W_{TSS+1}$), where TSS is the transcription start site, $N_{TSS-1}N_{TSS}$ is the primer binding site, Y is pyrimidine, and W is A or T. The same or essentially the same consensus is observed for all 16 primer sequences and for each major primer binding-site position (Figure S5). Analysis of the same promoter library *in vitro*, assessing primer-dependent initiation with the dinucleotide primer UpA,

yields the same consensus sequence, $Y_{TSS-2}N_{TSS-1}N_{TSS}W_{TSS+1}$, and does so for each primer binding-site position (Figures 4, right, and S6).

In vitro transcription experiments assessing competition between primer-dependent transcription initiation with UpA and primer-independent transcription initiation with ATP show that primer-dependent initiation is ~60 times more efficient than primer-independent initiation at a promoter conforming to the consensus sequence ($T_{TSS-2}T_{TSS-1}A_{TSS}T_{TSS+1}$; Figure 5A), but is only ~10 times more efficient at a promoter not having the consensus sequence ($G_{TSS-2}T_{TSS-1}A_{TSS}T_{TSS+1}$; Figure 5A).

In vitro transcription experiments using heteroduplex templates and the dinucleotide primer UpA show that the sequence information responsible for the preference for Y:R at TSS-2 resides exclusively in the DNA template strand (Figure 5B). Thus, in experiments with heteroduplex templates, primer-dependent initiation is reduced by replacement of the consensus nucleotide by a non-consensus nucleotide or an abasic site on the DNA template strand, but is not reduced by replacement of the consensus nucleotide by a non-consensus nucleotide or an abasic site on the DNA nontemplate strand (Figure 5B).

We conclude that primer-dependent initiation, *in vivo* and *in vitro*, depends not only on the sequence of the primer binding site, but also on flanking sequences, with the preferred sequence being $Y_{TSS-2}N_{TSS-1}N_{TSS}W_{TSS+1}$ ($Y{:}R_{TSS-2}N{:}N_{TSS-1}N{:}N_{TSS}W{:}W_{TSS+1}$).

Primer-dependent initiation: chromosomal promoters

To assess whether the sequence preferences observed in the MASTER analysis apply also to natural promoters, we quantified primer-dependent initiation in stationary-phase *E. coli* cells at each of 93 promoters that use UpA as primer (Table S1, Figure 6A). The results show the same sequence preferences at positions TSS-2 and TSS+1 observed in the MASTER analysis are observed in chromosomally-encoded promoters (Figure 6A).

To assess directly the functional significance of the sequence preferences observed in the MASTER analysis and natural promoter analysis, we constructed mutations at positions TSS-2 and TSS+1 of a natural promoter that uses UpA as primer (Figure 6B, top) and assessed effects on function in stationary-phase *E. coli* cells (Figure 6B, bottom). We observed that, at position TSS-2, the consensus base pair T:A is preferred over the non-consensus base pair G:C by a factor of ~5 (Figure 6B, bottom), and, at position TSS+1, the consensus base pair T:A is preferred over the non-consensus base pair G:C by a factor of ~2.5 (Figure 6B, bottom). We conclude that the sequence dependence for primer-dependent initiation defined using MASTER is also observed in natural, chromosomally-encoded *E. coli* promoters.

Primer-dependent initiation: structural basis of promoter sequence-dependence

To determine the structural basis of the preference for a template-strand purine nucleotide at position TSS-2 (R_{TSS-2}) in primer-dependent initiation, we determined crystal structures of transcription initiation complexes containing a template-strand purine nucleotide at position TSS-2 (Figure 7). We first prepared crystals of *Thermus thermophilus* RPo using synthetic nucleotide scaffolds containing a template-strand purine nucleotide, A, at position TSS-2, and containing a template-strand primer binding site for either the dinucleotide primer used most frequently in primer-dependent initiation *in vivo*, UpA (9, 10; Figure 3A), or the dinucleotide primer used second most frequently in primer-dependent initiation *in vivo*, GpG (9, 10; Figure 3A). We next soaked the crystals either with UpA and CMPcPP or with GpG and CMPcPP, to yield crystals of *T. thermophilus* RPo in complex with a dinucleotide primer and a non-reactive analog of an extending NTP. We then collected X-ray diffraction data, solved structures, and refined structures, obtaining structures of RPo[$A_{TSS-2}A_{TSS-1}T_{TSS}$]-UpA-CMPcPP at 2.8 Å resolution and RPo[$A_{TSS-2}C_{TSS-1}C_{TSS}$]-GpG-CMPcPP at 2.9 Å resolution (Table S2, Figure 7).

For both RPo[$A_{TSS-2}A_{TSS-1}T_{TSS}$]-UpA-CMPcPP and RPo[$A_{TSS-2}C_{TSS-1}C_{TSS}$]-GpG-CMPcPP, experimental electron density maps show unambiguous density for the dinucleotide primer in the RNAP active-center P-1 and P sites and for CMPcPP in the RNAP active-center A site (Figure 7B). The dinucleotide primers make extensive interactions with RNAP, template-strand DNA, and CMPcPP. For each dinucleotide primer, the primer phosphate, makes the same interactions with RNAP (residues βH1237, βK1065 and βK1073; residues numbered as in *E. coli* RNAP) and the RNAP active-center catalytic Mg^{2+} as the primer phosphate in a previously reported structure of *T. thermophilus* RPo-GpA (18; Figure 7A-C). For each dinucleotide primer, the primer bases make Watson-Crick H-bonds with template-strand nucleotides at TSS-1 and TSS, and make an intra-chain stacking interaction with the base of CMPcPP. Crucially, for each dinucleotide primer, the 5′ base of the primer makes an inter-chain stacking interaction with the base of the template-strand purine nucleotide at position TSS-2 (Figure 7). Structural modeling indicates that this inter-chain base-stacking interaction can occur only when the template-strand nucleotide at position TSS-2 is a purine (Figure 7D). Structural modeling further indicates that this inter-chain base-stacking interaction should stabilize the binding of the primer and CMPcPP to template-strand DNA, thereby facilitating primer-dependent initiation. Consistent with this structural modeling, a crystal structure of a complex obtained by soaking crystals of *T. thermophilus* RPo containing a template-strand pyrimidine nucleotide, T, at position TSS-2 with GpG and CMPcPP (RPo[$T_{TSS-2}C_{TSS-1}C_{TSS}$]-GpG; 3.4 Å resolution; Table S2) do not show significant inter-chain base-stacking with the template-strand at position TSS-2, and do not show binding of CMPcPP to template-strand DNA (Figure S7). Taken together, the crystal structures in Figures 7 and S7 define the structural basis of the preference for purine vs. pyrimidine at template-strand position TSS-2 in

primer-dependent transcription initiation: namely, inter-chain base stacking between the primer 5′ base and a purine at template-strand position TSS-2 facilitates binding of the primer and the extending NTP.

Discussion

Promoter-sequence dependence of primer-dependent initiation: mechanism and structural basis

Our biochemical results show that primer-dependent initiation in stationary-phase *E. coli* almost always involves a dinucleotide primer (Figures 2, S1), can involve any of the 16 possible dinucleotide primers (Figure 3A), almost always involves a primer binding site complementary to both the 5′ and 3′ nucleotides of the dinucleotide primer (Figures 3B, S2-S4), depends on promoter sequences flanking the primer binding site (Figures 4-6, S5-S6), and exhibits the consensus sequence $Y_{TSS-2}N_{TSS-1}N_{TSS}W_{TSS+1}$ ($Y{:}R_{TSS-2}N{:}N_{TSS-1}N{:}N_{TSS}W{:}W_{TSS+1}$; Figures 4-6, S5-S6), wherein the sequence information at positions TSS-2, TSS-1, and TSS is contained exclusively within the promoter template strand (Figures 5, S4).

Our structural results show that the sequence preference for purine at template-strand position TSS-2 is a consequence of inter-chain base stacking of a purine at template-strand position TSS-2 with the 5′ nucleotide of a dinucleotide primer (Figures 7, S7). The structural basis of the preference for purine vs. pyrimidine at template-strand position TSS-2 in primer-dependent initiation is analogous to--almost identical to--the previously described structural basis of the preference for purine vs. pyrimidine at template-strand position TSS-1 in primer-independent initiation (19, 20). In the former case, inter-chain base stacking between the primer 5′ base in the RNAP active-center P-1 site and a purine at template-strand position TSS-2 facilitates binding of the primer and an extending NTP. In the latter case, inter-chain base-stacking between the initiating NTP in the RNAP active-center P site and a purine at template-strand position TSS-1 facilitates binding of the initiating NTP and an extending NTP.

Promoter sequences upstream of the TSS modulate the chemical nature of the RNA 5′ end

Chemical modifications of the RNA 5′ end provide a layer of "epitranscriptomic" regulation, influencing multiple aspects of RNA fate, including stability, processing, localization, and translation efficiency (21-24). Primer-dependent initiation provides one mechanism to alter the RNA 5′ end during transcription initiation. In primer-dependent initiation with a dinucleotide primer, the RNA product acquires a 5′ hydroxyl and acquires one additional nucleotide at the RNA 5′ end (Figure 1). In an analogous manner, NCIN-dependent initiation--where an NCIN is a non-canonical initiating nucleotide--provides another mechanism to alter the RNA 5′ end during transcription initiation (25, 26). In NCIN-dependent initiation, the RNA product acquires an NCIN at the RNA 5′ end. NCIN-dependent

initiation has been shown to occur with the oxidized and reduced forms of nicotinamide adenine dinucleotide (NAD), dephospho-coenzyme A (dpCoA), flavin adenine dinucleotide (FAD), uridine diphosphate N-acetylglucosamine (UDP-GlcNAc), and dinucleoside tetraphosphates (Np4Ns) (25, 27-30)

All three modes of transcription initiation--primer-dependent, NCIN-dependent, and NTP-dependent--exhibit promoter-sequence dependence (12, 16, 25, 30; Figures 3-6, S2-S6). All three modes of transcription initiation exhibit promoter consensus sequences that include base pairs upstream of the initiating-entity binding site (12, 16, 25, 30; Figures 4-6, S5-S6). Crucially, the promoter consensus sequences upstream of the initiating-entity binding site for primer-dependent, NCIN-dependent (NAD, dpCoA, and Np4N), and NTP-dependent transcription initiation all are different: $Y:R_{TSS-2}$, $R:Y_{TSS-1}$, and $Y:R_{TSS-1}$, respectively (12, 16, 25, 30; Figures 4-6, S5-S6). It follows that the sequence of the promoter TSS region hard-codes not only the TSS position, but also the relative efficiencies of, and potentials for epitranscriptomic regulation through, primer-dependent, NCIN-dependent, and NTP-dependent transcription initiation.

Materials and Methods

Proteins

E. coli RNAP core enzyme used in transcription experiments was prepared from *E. coli* strain NiCo21(DE3) (New England Biolabs, NEB) transformed with plasmid pIA900 (31) using culture and induction procedures, immobilized-metal-ion affinity chromatography on Ni-NTA agarose, and affinity chromatography on Heparin HP as described in (32). *E. coli* σ^{70} was prepared from *E. coli* strain NiCo21 (DE3) transformed with plasmid pσ^{70}-His using culture and induction procedures, immobilized-metal-ion affinity chromatography on Ni-NTA agarose, and anion-exchange chromatography on Mono Q as described in (33). 10x RNAP holoenzyme was formed by mixing 0.5 μM RNAP core and 2.5 μM σ^{70} in 1x reaction buffer (40 mM Tris HCl, pH 7.5; 10 mM $MgCl_2$; 150 mM KCl; 0.01% Triton X-100; and 1 mM DTT).

5' RNA polyphosphatase (Rpp) and T4 Polynucleotide Kinase (PNK) were purchased from Epicentre and NEB, respectively.

Oligonucleotides

Oligodeoxyribonucleotides (Table S3) were purchased from Integrated DNA Technologies (IDT) and were purified with standard desalting purification. UpA and GpG were purchased from Trilink Biotechnologies. NTPs (ATP, GTP, CTP, and UTP) were purchased from GE Healthcare Life Sciences.

Homoduplex and heteroduplex templates used in single-template *in vitro* transcription assays were generated by mixing 1.1 μM nontemplate-strand oligo with 1 μM template-strand oligo in 10 mM Tris (pH 8.0). Mixtures were heated to 90°C for 10 min and slowly cooled to 40°C (0.1°C / second) using a Dyad PCR machine (Bio-Rad).

Plasmids

Plasmid pBEN516 (9) contains sequences from -100 to +15 of the *bhsA* promoter fused to the tR' terminator inserted between the HindIII and SalI sites of pACYC184 (NEB). Mutant derivatives of pBEN516 containing a G:C base pair at position TSS-2 (pKS494) or a G:C base pair at position +2 (pKS497) were generated using site-directed mutagenesis. Plasmid pPSV38 (9) contains a pBR322 origin, a gentamycin resistance gene (*aaC1*), and *lacI*q. Plasmid library pMASTER-*lac*CONS-N10 has been previously described (12).

Analysis of primer-dependent initiation by MASTER (Figures 2-4, S1-S3, S5-S6)

Primer-dependent initiation in vitro: transcription reaction conditions

A linear DNA fragment containing p*lac*CONS-N10 generated as described in (34) was used as template for *in vitro* transcription assays. Transcription reactions (total volume = 100 µl) were performed by mixing 10 nM of template DNA with 50 nM RNAP holoenzyme in *E. coli* RNA polymerase reaction buffer (NEB) (40 mM Tris HCl, pH 7.5; 10 mM $MgCl_2$; 150 mM KCl; 0.01% Triton X-100; 1 mM DTT), 0.1 mg/ml BSA (NEB), and 40 U murine RNase inhibitor (NEB). Reactions were incubated at 37°C for 15 min to form open complexes. A single round of transcription was initiated by addition of 1000 µM ATP, 1000 µM CTP, 1000 µM UTP, 1000 µM GTP, UpA (40 µM, 160 µM, or 640 µM), and 0.1 mg/ml heparin (Sigma Aldrich). Reactions were incubated at 37°C for 15 min and stopped by addition of 0.5 M EDTA (pH 8) to a final concentration of 50 mM. (For each replicate, two 100 µl transcription reactions were performed separately and combined after addition of EDTA.) Nucleic acids were recovered by ethanol precipitation, reconstituted in 25 µl of nuclease-free water, mixed with 25 µl of 2x RNA loading dye (95% formamide; 25 mM EDTA; 0.025% SDS; 0.025% xylene cyanol; 0.025% bromophenol blue), and separated by electrophoresis on 10% 7M urea slab gels (Invitrogen) equilibrated and run in 1x TBE. The gel was stained with SYBR Gold nucleic acid gel stain (Invitrogen), bands visualized on a UV transilluminator, and RNA products ~150 nt in length were excised from the gel. The excised gel slice was crushed, 300 µl of 0.3 M NaCl in 1x TE buffer was added, and the mixture was incubated at 70°C for 10 min. Eluted RNAs were collected using a Spin-X column (Corning). After the first elution, the crushed gel fragments were collected and the elution procedure was repeated, nucleic acids were collected, pooled with the first elution, isolated by isopropanol precipitation, and resuspended in 25.5 µl of RNase-free water (Invitrogen). Reactions were performed in triplicate.

Primer-dependent initiation in stationary-phase E. coli cells: cell growth

Three independent 25 ml cell cultures of *E. coli* MG1655 cells (gift of A. Hochschild, Harvard Medical School) containing p*lac*CONS-N10 and pPSV38 were grown in LB media (Millipore) containing chloramphenicol (25 µg/ml), gentamicin (10 µg/ml), and IPTG (1 mM) in a 125 ml DeLong flask (Bellco Glass) shaken at 220 RPM at 37°C until late stationary phase (~21 hours after entry into stationary phase; final OD600 ~3.5). 2 ml aliquots of cell suspensions were placed in 2 ml tubes and cells were collected by centrifugation (1 min; 21,000 x *g*; 20°C). Supernatants were removed and cells stored at -80°C.

Primer-dependent initiation in stationary-phase E. coli cells: RNA isolation

RNA was isolated from frozen cell pellets as described in (12). Cell pellets were resuspended in 600 µl of TRI Reagent solution (Molecular Research Center), incubated at 70°C for 10 min, and centrifuged (10 min; 21,000 x *g*; 4°C) to remove insoluble material. The supernatant was transferred to a fresh tube, ethanol was added to a final concentration of 60.5%, and the mixture was applied to a

Direct-zol spin column (Zymo Research). DNase I (Zymo Research) treatment was performed on-column according to the manufacturer's recommendations. RNA was eluted from the column using nuclease-free water heated to 70°C (3 x 30 µl elutions; total volume of eluate = 90 µl). RNA was treated with 2 U TURBO DNase (Invitrogen) at 37°C for 1 h, samples were extracted with acid phenol:chloroform (Ambion), RNA was recovered by ethanol precipitation and resuspended in RNase-free water. A MICROBExpress Kit (Invitrogen) was used to remove rRNAs from ~36 µg of recovered RNA, rRNA-depleted RNA was isolated by ethanol precipitation and resuspended in 40 µl of RNase-free water.

Enzymatic treatment of RNA products

For RNAs isolated from *E. coli*, 3 µg of rRNA-depleted RNA was used in each reaction. RNAs isolated from *in vitro* transcription reactions were split into four equal portions and used in each reaction.

Rpp treatment (total reaction volume = 30 µl): RNA products were mixed with 20 U Rpp and 40 U RNaseOUT (Invitrogen) in 1x Rpp reaction buffer (50 mM HEPES-KOH, pH 7.5; 100 mM NaCl; 1 mM EDTA; 0.1% BME; and 0.01% Triton X-100) and incubated at 37°C for 1 hr. Reactions were extracted with acid phenol:chloroform, RNA was recovered by ethanol precipitation, and resuspended in 10.5 µl RNase-free water.

PNK treatment (total reaction volume = 50 µl): RNA products were mixed with 20 U PNK, 40 U RNaseOUT, and 1 mM ATP (NEB) in 1x PNK reaction buffer (70 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 5 mM DTT) and incubated at 37°C for 1 hr. Processed RNAs were recovered using Qiagen's RNeasy MinElute kit (following the manufacturer's recommendations with the exception that RNAs were eluted from the column using 200 µl nuclease-free water heated to 70°C). RNA was recovered by ethanol precipitation and resuspended in 10.5 µl RNase-free water.

Rpp and PNK treatment: RNA products were mixed with 20 U PNK, 40 U RNaseOUT, and 1 mM ATP in 1x PNK reaction buffer (total reaction volume = 50 µl) and incubated at 37°C for 1 hr. Processed RNAs were recovered using Qiagen's RNeasy MinElute kit (following the manufacturer's recommendations with the exception that RNAs were eluted from the column using 25 µL nuclease-free water heated to 70°C). Recovered RNA products were mixed with 20 U Rpp and 40 U RNaseOUT in 1x Rpp reaction buffer (total reaction volume = 30 µl) and incubated at 37°C for 1 hr. Reactions were extracted with acid phenol:chloroform, RNA was recovered by ethanol precipitation, and resuspended in 10.5 µl RNase-free water.

"mock" PNK treatment (total reaction volume = 50 µl): RNA products were mixed with 40 U RNaseOUT and 1 mM ATP in 1x PNK reaction buffer and incubated at 37°C for 1 hr. Reactions were

extracted with acid phenol:chloroform, RNA was recovered by ethanol precipitation, and resuspended in 10.5 µl RNase-free water.

"mock" Rpp treatment (total reaction volume = 30 µl): RNA products were mixed with 40 U RNaseOUT in 1x Rpp reaction buffer (total reaction volume = 30 µl) and incubated at 37°C for 1 hr. Reactions were extracted with acid phenol:chloroform, RNA was recovered by ethanol precipitation, and resuspended in 10.5 µl RNase-free water.

5'-adaptor ligation

To enable quantitative comparisons between samples, we performed the 5'-adaptor ligation step using barcoded 5'-adaptor oligonucleotides as described in (16). For RNA products isolated from stationary-phase *E. coli* cells, oligo i105 was used for RNAs processed by Rpp, oligo i106 was used for RNAs processed with PNK, oligo i107 was used for RNAs processed with both Rpp and PNK, and oligo i108 was used for unprocessed RNAs (mock PNK treated). For RNA products isolated from *in vitro* reactions, oligo i105 was used for RNAs processed by Rpp, oligo i106 was used for unprocessed RNAs (mock Rpp treated); oligo i107 was used for RNAs processed by PNK, and oligo i108 was used for unprocessed RNAs (mock PNK treated).

Processed RNA products isolated from stationary-phase *E. coli* cells (in 10.5 µl of nuclease-free water) were combined with 1 mM ATP (NEB), 40 U RNaseOUT, 1x T4 RNA ligase buffer (NEB), and 10 U T4 RNA ligase 1 (NEB) and 1 µM of 5' adaptor oligo (total reaction volume = 20 µl), and incubated at 37°C for 2 h. Reactions were then supplemented with 1x T4 RNA ligase buffer, 1 mM ATP, PEG 8000 (10% final), 5U T4 RNA ligase 1, and 20 U RNaseOUT (total reaction volume = 30 µl) and further incubated at 16°C for 16 h. Processed RNA products isolated from *in vitro* reactions (in 10.5 µl of nuclease-free water) were combined with PEG 8000 (10% final concentration), 1 mM ATP, 40 U RNaseOUT, 1x T4 RNA ligase buffer, 10 U T4 RNA ligase 1, and 1 µM of 5' adaptor oligo (total reaction volume = 30 µl), and incubated at 16°C for 16 h.

Ligation reactions were stopped by addition of 30 µl of 2x RNA loading dye and heated at 95°C for 5 min. For each replicate, the 4 ligation reactions were combined, and separated by electrophoresis on 10% 7M urea slab gels (equilibrated and run in 1x TBE). Gels were incubated with SYBR Gold nucleic acid gel stain, and bands were visualized with UV transillumination. For RNAs isolated from stationary-phase *E. coli* cells, products migrating above the 5'-adapter oligo were isolated from the gel (procedures as above; recovered in 50 µl of nuclease-free water). For RNAs generated *in vitro*, products ~150 nt in length were recovered from the gel (procedures as above; recovered in 16 µl of nuclease-free water). 5'-adaptor-ligated RNAs were used for analysis of primer-dependent initiation from p*lac*CONS-10 (this section) and for analysis of primer-dependent initiation from natural, chromosomally-encoded

promoters (next section).

First strand cDNA synthesis

For RNAs isolated from stationary-phase *E. coli* cells, 25 µl of 5'-adaptor-ligated RNAs were mixed with 1.5 µl s128A oligonucleotide (3 µM) and 3.5 µl nuclease-free water. The 30 µl mixture was incubated at 65°C for 5 min, cooled to 4°C, and combined with 20 µl of a solution containing 10 µl of 5x First-Strand buffer (Invitrogen), 2.5 µl of 10 mM dNTP mix (NEB), 2.5 µl of 100 mM DTT (Invitrogen), 2.5 µl 40 U/µl RNaseOUT, and 2.5 µl 100 U/µl SuperScript III Reverse Transcriptase (Invitrogen), for a final reaction volume of 50 µl. Reactions were incubated at 25°C for 5 min, 55°C for 60 min, 70°C for 15 min, then kept at 25°C. Next, 5.4 µl of 1M NaOH was added, reactions were incubated at 95°C for 5 min, and kept at 10°C, 4.5 µl 1.2M HCl was added, followed by 60 µl of 2x RNA loading dye.

For RNAs isolated from *in vitro* reactions, 16 µl of 5'-adaptor-ligated RNAs were mixed with 0.5 µl s128A oligonucleotide (1.5 µM). The 16.5 µl mixture was incubated at 65°C for 5 min, cooled to 4°C, and combined with 13.5 µl of a solution containing 6 µl of 5x First-Strand buffer, 1.5 µl of 10 mM dNTP mix, 1.5 µl of 100 mM DTT, 1 µl 40 U/µl RNaseOUT, 1.5 µl 100 U/µl SuperScript III Reverse Transcriptase, and 2 µl of nuclease-free water, for a final reaction volume of 30 µl. Reactions were incubated at 25°C for 5 minutes, 55°C for 60 minutes, 70°C for 15 min, then cooled to 25°C. Next, 10 U of RNase H (NEB) was added, reactions were incubated at 37°C for 15 min, and 31 µl of 2x RNA loading dye was added.

Nucleic acids were separated by electrophoresis on 10% 7M urea slab gels (equilibrated and run in 1x TBE). Gels were incubated with SYBR Gold nucleic acid gel stain, bands were visualized with UV transillumination, and species ~80 to ~150 nt in length were recovered from the gel (procedure as above) and recovered in 20 µl of nuclease-free water.

cDNA amplification

cDNA derived from RNA products generated *in vitro* or *in vivo* were diluted with nuclease-free water to a concentration of ~10^9 molecules/µl. 2µl of the diluted cDNA solution was used as a template for emulsion PCR reactions containing Illumina index primers using a Micellula DNA Emulsion and Purification Kit (EURx). The Illumina PCR forward primer and Illumina index primers from the TruSeq Small RNA Sample Prep Kits were used. The emulsion was broken, and DNA was purified according to the manufacturer's recommendations. Amplicons were gel purified on 10% TBE slab gels (Invitrogen; equilibrated in 1x TBE), recovered by isopropanol precipitation and reconstituted in 13 µl of nuclease-free water.

High-throughput sequencing.

Barcoded libraries were pooled and sequenced on an Illumina NextSeq platform in high-output mode using custom sequencing primer s1115.

Sample serial numbers

Samples KS112-KS114 are cDNA derived from RNA products generated in stationary-phase *E. coli* cells treated with (i) both PNK and Rpp (PNK + Rpp), (ii) Rpp only (Rpp), (iii) PNK only (PNK), or (iv) neither PNK nor Rpp (mock). Samples KS86-KS97 are cDNA derived from RNA products generated *in vitro* in the presence of no UpA (KS86-KS88), 40 μM UpA (KS89-KS91), 160 μM UpA (KS92-KS94), or 640 μM UpA (KS95-KS97) treated with (i) Rpp only (Rpp), (ii) PNK only (PNK), or (iii) neither PNK nor Rpp (mock).

Data analysis: separation of RNA 5'-end sequences by enzymatic treatment, promoter sequence, and promoter position

RNA 5'-end sequences were associated with an enzymatic treatment using the 4-nt barcode sequence acquired upon ligation of the 5'-adaptor (see above) as described in (16). RNA 5'-end sequences were associated with a p*lac*CONS promoter sequence using transcribed-region barcode assignments derived from the analysis of sample Vv945 described in (14). RNA 5'-end sequences that could be aligned to their template of origin with no mismatches were used for results presented in Figures 2, 3A, 4, S1, S5, and S6. RNA 5'-end sequences with mismatches at the first and/or second base of the 5'-end were also included for results shown in Figures 3B, S2-S3.

Data analysis: 5'-end distribution histograms (Figures 2, S1B, S8)

The number of 5'-end sequences emanating from each position 4 to 10 bp downstream of the -10 element of p*lac*CONS was determined for each of the ~4^{10} (~1,000,000) promoter sequences. These counts are represented using four vectors, $\vec{c}_{\mathrm{PNK+Rpp}}$, $\vec{c}_{\mathrm{Rpp}}$, $\vec{c}_{\mathrm{PNK}}$, and $\vec{c}_{\mathrm{mock}}$, which represent the number of counts observed for 5'-ends at positions $i = 4, 5, \ldots, 10$ for each enzymatic treatment. We initially estimated the number of 5'-ppp RNAs and 5'-OH RNAs in two different ways:

We initially computed the number of 5'-ppp RNAs and 5'-OH RNAs in two different ways:

$$\vec{c}_{\mathrm{ppp1}} = \vec{c}_{\mathrm{PNK+Rpp}} - \vec{c}_{\mathrm{PNK}},$$
$$\vec{c}_{\mathrm{ppp2}} = \vec{c}_{\mathrm{Rpp}} - \vec{c}_{\mathrm{mock}},$$
$$\vec{c}_{\mathrm{OH1}} = \vec{c}_{\mathrm{PNK+Rpp}} - \vec{c}_{\mathrm{Rpp}},$$
$$\vec{c}_{\mathrm{OH2}} = \vec{c}_{\mathrm{PNK}} - \vec{c}_{\mathrm{mock}}.$$

These four read count distributions were computed separately for each of the three replicates. To visualize these distributions, we normalized each counts vector by its sum across all positions, i.e.,

$$\vec{p}_{\mathrm{X}}(i) = \frac{\vec{c}_{\mathrm{X}}(i)}{\sum_j \vec{c}_{\mathrm{X}}(j)}, \qquad X \in \{\mathrm{ppp1}, \mathrm{ppp2}, \mathrm{OH1}, \mathrm{OH2}\}.$$

The 5'-OH distributions that resulted exhibited two obvious problems (Figure S8A, left): there was substantial variation across replicates, and four of the distributions exhibited probabilities well below zero at two positions ($i = 8{,}9$). We reasoned that these defects might be artefacts resulting from the enzymatic treatments not being 100% efficient, and that accounting for these inefficiencies might lead to more accurate 5'-OH distributions. Let $\epsilon_{\mathrm{PNK+Rpp}}$, ϵ_{PNK}, ϵ_{Rpp}, and ϵ_{mock} denote the efficiencies of the four enzymatic treatments. Then the number of true underlying counts in the four samples becomes,

$$\vec{c}_{\mathrm{ppp1}} = \frac{\vec{c}_{\mathrm{PNK+Rpp}}}{\epsilon_{\mathrm{PNK+Rpp}}} - \frac{\vec{c}_{\mathrm{PNK}}}{\epsilon_{\mathrm{PNK}}},$$

$$\vec{c}_{\mathrm{ppp2}} = \frac{\vec{c}_{\mathrm{Rpp}}}{\epsilon_{\mathrm{Rpp}}} - \frac{\vec{c}_{\mathrm{mock}}}{\epsilon_{\mathrm{mock}}},$$

$$\vec{c}_{\mathrm{OH1}} = \frac{\vec{c}_{\mathrm{PNK+Rpp}}}{\epsilon_{\mathrm{PNK+Rpp}}} - \frac{\vec{c}_{\mathrm{Rpp}}}{\epsilon_{\mathrm{Rpp}}},$$

$$\vec{c}_{\mathrm{OH2}} = \frac{\vec{c}_{\mathrm{PNK}}}{\epsilon_{\mathrm{PNK}}} - \frac{\vec{c}_{\mathrm{mock}}}{\epsilon_{\mathrm{mock}}}.$$

One can solve for the unknown efficiencies by setting $\vec{c}_{\mathrm{ppp1}} = \vec{c}_{\mathrm{ppp2}}$, or equivalently $\vec{c}_{\mathrm{OH1}} = \vec{c}_{\mathrm{OH2}}$, both of which give

$$\frac{\vec{c}_{\mathrm{PNK+Rpp}}}{\epsilon_{\mathrm{PNK+Rpp}}} - \frac{\vec{c}_{\mathrm{PNK}}}{\epsilon_{\mathrm{PNK}}} - \frac{\vec{c}_{\mathrm{Rpp}}}{\epsilon_{\mathrm{Rpp}}} + \frac{\vec{c}_{\mathrm{mock}}}{\epsilon_{\mathrm{mock}}} = 0.$$

This provides a system of 7 equations with 4 unknowns, and setting $\epsilon_{\mathrm{mock}} = 1$ allowed us to solve for the other three other (now relative) efficiencies. We did this by minimizing the objective function

$$\sum_i \frac{\vec{r}(i)^2}{\vec{v}(i)},$$

where

$$\vec{r} = \frac{\vec{c}_{\mathrm{PNK+Rpp}}}{\epsilon_{\mathrm{PNK+Rpp}}} - \frac{\vec{c}_{\mathrm{PNK}}}{\epsilon_{\mathrm{PNK}}} - \frac{\vec{c}_{\mathrm{Rpp}}}{\epsilon_{\mathrm{Rpp}}} + \frac{\vec{c}_{\mathrm{mock}}}{\epsilon_{\mathrm{mock}}},$$

$$\vec{v} = \frac{\vec{c}_{\mathrm{PNK+Rpp}}}{\epsilon^2_{\mathrm{PNK+Rpp}}} + \frac{\vec{c}_{\mathrm{PNK}}}{\epsilon^2_{\mathrm{PNK}}} + \frac{\vec{c}_{\mathrm{Rpp}}}{\epsilon^2_{\mathrm{Rpp}}} + \frac{\vec{c}_{\mathrm{mock}}}{\epsilon^2_{\mathrm{mock}}},$$

are vectors that respectively represent the residuals and Poisson-estimated variances (Figure S8B). Using these efficiencies, we computed the corrected read count distributions (Figure S8A, right). The resulting 5'-OH distributions were much more reproducible across replicates. Moreover, a negative probability was estimated only for position 9 for $\vec{c}_{\mathrm{OH,2}}$ of replicate 2, and even this was much closer to zero than in the

uncorrected profiles (Figure S8A, right). We therefore chose to compute 5'-ppp and 5'-OH read counts using the averages,

$$\vec{c}_{\mathrm{PPP}} = \frac{1}{2}\left(\frac{\vec{c}_{\mathrm{PNK+Rpp}}}{\epsilon_{\mathrm{PNK+Rpp}}} - \frac{\vec{c}_{\mathrm{PNK}}}{\epsilon_{\mathrm{PNK}}} + \frac{\vec{c}_{\mathrm{Rpp}}}{\epsilon_{\mathrm{Rpp}}} - \frac{\vec{c}_{\mathrm{mock}}}{\epsilon_{\mathrm{mock}}}\right),$$

$$\vec{c}_{\mathrm{OH}} = \frac{1}{2}\left(\frac{\vec{c}_{\mathrm{PNK+Rpp}}}{\epsilon_{\mathrm{PNK+Rpp}}} + \frac{\vec{c}_{\mathrm{PNK}}}{\epsilon_{\mathrm{PNK}}} - \frac{\vec{c}_{\mathrm{Rpp}}}{\epsilon_{\mathrm{Rpp}}} - \frac{\vec{c}_{\mathrm{mock}}}{\epsilon_{\mathrm{mock}}}\right).$$

These averaged distributions were used in the *in vivo* MASTER analysis shown in Figures 2 and S1A, as were analogous formulas for analyzing primer usage and for generating the sequence logos shown in Figures 3-4 and S5.

Data analysis: modeling 5'-OH distributions as a mixture of shifted 5'-ppp distributions (Figure S1A)

We modeled the 5'-end distributions of 5'-OH RNAs as a mixture of 5'-ppp RNAs, each shifted between -4 nt and +4 nt. The mixture coefficients were inferred using least-squares regression under positivity and normalization constraints. The 5'-ppp distributions shifted one nucleotide upstream account for 78.7% ± 4.8% of the mixture model, no shift in the 5'-ppp distributions accounts for 20.0% ± 4.8% of the mixture model, and all other shifts account for 0.39% ± 0.06% of the model (uncertainties represent the SD across replicates).

To carry out the mixture modeling of 5'-end distributions, we defined a 7x9 matrix A, whose entries A_{ij} represent the fraction of reads with 5'-ends at positions 4-10 (corresponding to $i = 1, 2, \ldots, 7$ within the 5'-ppp distribution shifted by -4, -3, ..., +4 nt (corresponding to $j = 1, 2, \ldots, 9$). Note that these shifted distributions were normalized so that $\sum_i A_{ij} = 1$ for every column j. We further defined a 7x1 vector $\vec{b}$ representing the 5'-OH distribution, normalized so that $\sum_i b_i = 1$. We then inferred a 9x1 vector of mixture coefficients $\vec{x}$ by solving the constrained least squares problem

$$\vec{x}^* = \mathrm{argmin} \left\| A\vec{x} - \vec{b} \right\|^2$$

under the constraints that all $x_j \geq 0$ and that $\sum_j x_j = 1$. The resulting mixture distribution $\vec{b}^* = A\vec{x}^*$ is shown in Figure S1A (left panel), alongside the 5'-OH distribution $\vec{b}$. The corresponding mixture coefficients are shown in Figure S1A (right panel). The residual deviation was computed as

$$\frac{1}{2}\sum_i |b_i^* - b_i|.$$

Data analysis: in vivo sequence logos (Figures 4, S5)

Sequence logos illustrate the sequence-dependent $\log_2$ likelihood of primer-dependent initiation for primer binding sites 6-7, 7-8, and 8-9. All logos were created using Logomaker (35). The specific quantities illustrated in the logos were computed as follows.

To generate the logos shown in Figures 4 (left) and S5**,** we first estimated the values of $\vec{c}_{\mathrm{ppp}}$ and $\vec{c}_{\mathrm{OH}}$ originating from promoters for which each base $b \in \{A, C, G, T\}$ occurs at each position $l \in \{1,2,\ldots,10\}$ within the 10 bp randomized region (i.e., positions 1-10 bp downstream of the -10 element). These read count estimates were computed as

$$\mathrm{c}_{b,l}^{\mathrm{ppp}}(\mathrm{i}) = \min\left\{0, \frac{1}{2}\left[\frac{\mathrm{c}_{b,l}^{\mathrm{PNK+Rpp}}(i)}{\epsilon_{\mathrm{PNK+Rpp}}} - \frac{\mathrm{c}_{b,l}^{\mathrm{PNK}}(i)}{\epsilon_{\mathrm{PNK}}} + \frac{\mathrm{c}_{b,l}^{\mathrm{Rpp}}(i)}{\epsilon_{\mathrm{Rpp}}} - \frac{\mathrm{c}_{b,l}^{\mathrm{mock}}(i)}{\epsilon_{\mathrm{mock}}}\right]\right\} + 1,$$

$$\mathrm{c}_{b,l}^{\mathrm{OH}}(\mathrm{i}) = \min\left\{0, \frac{1}{2}\left[\frac{\mathrm{c}_{b,l}^{\mathrm{PNK+Rpp}}(i)}{\epsilon_{\mathrm{PNK+Rpp}}} + \frac{\mathrm{c}_{b,l}^{\mathrm{PNK}}(i)}{\epsilon_{\mathrm{PNK}}} - \frac{\mathrm{c}_{b,l}^{\mathrm{Rpp}}(i)}{\epsilon_{\mathrm{Rpp}}} - \frac{\mathrm{c}_{b,l}^{\mathrm{mock}}(i)}{\epsilon_{\mathrm{mock}}}\right]\right\} + 1,$$

where $i \in \{4, 5, \ldots, 10\}$ indicates the location of the 5' end of the tallied transcripts, and $c_{b,l}^{X}(i)$ is the number of transcripts observed to originate at position i from promoters with base b at position l in the sample corresponding to enzymatic treatment X. The minimum is to prevent counts from becoming negative, and the "+ 1" is a pseudo-count used to regularize the computation of log ratios. We then computed the total number of such reads summed over 5'-end positions i:

$$c_{b,l}^{\mathrm{total}} = \sum_{i=4}^{9}\left[\mathrm{c}_{b,l}^{\mathrm{OH}}(\mathrm{i}) + \mathrm{c}_{b,l}^{\mathrm{ppp}}(\mathrm{i}+1)\right].$$

A logo for each 5'-end position i was then plotted, where the height of each character b at each position l was given by the centered $\log_2$ ratio

$$h_{b,l}(i) = \log_2 \frac{\mathrm{c}_{b,l}^{\mathrm{OH}}(\mathrm{i})}{c_{b,l}^{\mathrm{total}}} - \bar{h}_l(i), \quad \bar{h}_l(i) = \frac{1}{4}\sum_b h_{b,l}(i).$$

Logos were computed separately for each of the three biological replicates. Logos shown in Figure 4 (left panel) were created using reads from all promoter sequences. Logos shown in Figures S5 and S6 were created using reads from promoter sequences with the specified primer binding site at positions i and $i+1$.

Data analysis: in vitro sequence logos (Figures 4, S6)

To generate the sequence logos shown in Figures 4 (right) and S6, we estimated the values of $\vec{c}_{\mathrm{ppp}}$ and $\vec{c}_{\mathrm{OH}}$ as

$$\mathrm{c}_{b,l}^{\mathrm{ppp}}(\mathrm{i}) = \min\left\{0, \mathrm{c}_{b,l}^{\mathrm{Rpp}}(i) - \mathrm{c}_{b,l}^{\mathrm{mockRpp}}(i)\right\} + 1,$$

$$\mathrm{c}_{b,l}^{\mathrm{OH}}(\mathrm{i}) = \min\left\{0, \mathrm{c}_{b,l}^{\mathrm{PNK}}(i) - \mathrm{c}_{b,l}^{\mathrm{mockPNK}}(i)\right\} + 1.$$

These estimated counts were used to generate logos for promoter sequences with a UpA binding site at positions i and $i+1$ using the procedure described above.

Data analysis: dinucleotide usage (Figure 3A)

The percent usage of each dinucleotide primer $x \in \{\mathrm{ApA}, \mathrm{ApC}, \ldots, \mathrm{UpU}\}$ was computed as

$$100 \times \frac{c_x^{\mathrm{OH}}(\mathrm{i})}{c_{\mathrm{total}}^{\mathrm{OH}}},$$

where

$$c_x^{\mathrm{OH}}(\mathrm{i}) = \min\left\{0, \frac{1}{2}\left[\frac{c_x^{\mathrm{PNK+Rpp}}(i)}{\epsilon_{\mathrm{PNK+Rpp}}} + \frac{c_x^{\mathrm{PNK}}(i)}{\epsilon_{\mathrm{PNK}}} - \frac{c_x^{\mathrm{Rpp}}(i)}{\epsilon_{\mathrm{Rpp}}} - \frac{c_x^{\mathrm{mock}}(i)}{\epsilon_{\mathrm{mock}}}\right]\right\} + 1$$

is the regularized efficiency-corrected count of 5'-OH RNAs originating from promoters with a binding site for primer x at positions i and $i+1$, and

$$c_{\mathrm{total}}^{\mathrm{OH}} = \sum_x \sum_i c_x^{\mathrm{OH}}(\mathrm{i})$$

is the sum of such counts over positions and primers.

Data analysis: primer binding site complementarity (Figures 3B, S2, S3)

To compute the binding site complementarity results, we first computed the efficiency-corrected regularized read counts

$$c_z^{\mathrm{OH}}(\mathrm{i}) = \min\left\{0, \frac{1}{2}\left[\frac{c_z^{\mathrm{PNK+Rpp}}(i)}{\epsilon_{\mathrm{PNK+Rpp}}} + \frac{c_z^{\mathrm{PNK}}(i)}{\epsilon_{\mathrm{PNK}}} - \frac{c_z^{\mathrm{Rpp}}(i)}{\epsilon_{\mathrm{Rpp}}} - \frac{c_z^{\mathrm{mock}}(i)}{\epsilon_{\mathrm{mock}}}\right]\right\} + 1,$$

which are conditioned on the indicator variable

$$z = \begin{cases} 11 & \text{match at TSS} - 1 \text{ and TSS} \\ 10 & \text{match at TSS} - 1, \text{mismatch at TSS} \\ 01 & \text{mismatch at TSS} - 1, \text{match at TSS} \\ 00 & \text{mismatches at TSS} - 1 \text{ and TSS} \end{cases}$$

The corresponding total counts were computed as

$$c_{\mathrm{all}}^{\mathrm{OH}}(i) = \sum_z c_z^{\mathrm{OH}}(\mathrm{i}).$$

The position-dependent "% 5'-OH RNA" values plotted in Figures S2 and S3 are given by

$$100 \times \frac{c_z^{\mathrm{OH}}(\mathrm{i})}{c_{\mathrm{all}}^{\mathrm{OH}}(\mathrm{i})}$$

for appropriate choice of z. The corresponding percentages aggregated across positions, shown in Figures 3B and S3, are given by

$$100 \times \frac{\sum_i c_z^{\mathrm{OH}}(\mathrm{i})}{\sum_i c_{\mathrm{all}}^{\mathrm{OH}}(\mathrm{i})},$$

where the sums are over positions 6, 7, and 8.

Analysis of primer-dependent initiation from chromosomally-encoded *E. coli* promoters (Figure 6A, Table S1)

cDNA library construction and sequencing

Cell growth, RNA isolation, enzymatic treatments, and 5'-adaptor ligations were performed as described above. 25 µl of 5'-adaptor-ligated RNAs were mixed with 5 µl of 18.5 µM (1.86 µM final) of an oligo pool consisting of a mixture of 93 gel purified oligodeoxyribonucleotides each having 5'-end sequence identical to the "RT primer" contained in Illumina TruSeq Small RNA Sample Prep Kits and a 3'-end sequence complementary to a chromosomally-encoded *E. coli* promoter that uses UpA as primer (Table S3). The mixture was incubated at 65°C for 5 min, kept at 4°C, combined with 20 µl of a solution containing 10 µl of 5x First-Strand buffer (Invitrogen), 2.5 µl of 10 mM dNTP mix (NEB), 2.5 µl of 100 mM DTT (Invitrogen), 2.5 µl of 40 U/µl RNaseOUT, and 2.5 µl of 100 U/µl SuperScript III Reverse Transcriptase (Invitrogen), for a final reaction volume of 50 µl. Reactions were incubated at 25°C for 5 min, 55°C for 60 min, 70°C for 15 min, then cooled to 25°C. Next, 5.4 µl of 1M NaOH was added, reactions were incubated at 95°C for 5 min, cooled to 10°C, 4.5 µl of 1.2M HCl was added, followed by 60 µl of 2x RNA loading dye. Nucleic acids were separated by electrophoresis on 10% 7M urea slab gels (equilibrated and run in 1x TBE). Gels were incubated with SYBR Gold nucleic acid gel stain, bands were visualized with UV transillumination, and species ~80 to ~150 nt in length were recovered from the gel (procedure as above) and recovered in 20 µl of nuclease-free water.

cDNA amplification and high-throughput sequencing was performed as described above. Serial numbers for these samples are KS118-KS120.

Data analysis: chromosomal promoter sequence logo (Figure 6A)

Sequencing reads were associated with one of the four reaction conditions based on the identity of the 4-nt barcode sequence. RNA 5'-end sequences that could be aligned to the chromosomally-encoded promoter from which they were expressed with no mismatches were used for results presented in Figure 6A. The number of 5'-end sequences emanating from each position up to four bases upstream and downstream of the UpA binding site (TSS-5 to TSS+4, where UpA binds positions TSS-1, TSS) was determined for each enzymatic treatment. To represent these data as a sequence logo, we first estimated the fraction of transcripts that had 5′-OH ends at TSS-1. For each promoter sequence *s*, this was computed using

$$r_s = \frac{c_s^{\mathrm{OH}}(\mathrm{TSS}-1)}{c_s^{\mathrm{total}}},$$

where

$$\mathrm{c}_s^{\mathrm{OH}}(\mathrm{i}) = \min\{0, \mathrm{c}_s^{\mathrm{PNK+Rpp}}(\mathrm{i}) - \mathrm{c}_s^{\mathrm{Rpp}}(\mathrm{i})\} + 1,$$

$$\mathrm{c}_s^{\mathrm{total}} = \sum_{i=\mathrm{TSS}-5}^{i=\mathrm{TSS}+4} \mathrm{c}_s^{\mathrm{PNK+Rpp}}(i) + L,$$

and where $\mathrm{c}_s^{\mathrm{PNK+Rpp}}(i)$ and $\mathrm{c}_s^{\mathrm{Rpp}}(\mathrm{i})$ denote the number of read initiating from position i (which ranges over L=10 positions, from TSS-5 to TSS+4) observed for promoter s in the PNK+Rpp and Rpp treatments, respectively. These r_s values were then averaged across three replicates, and a sequence logo reflecting the average $\log_2$ value of these ratios was rendered using mean-centered character heights given by

$$h_{b,l} = \frac{\sum_s s_{b,l} \log_2 r_s}{\sum_s s_{b,l}} - \bar{h}_l, \quad \bar{h}_l = \frac{1}{4}\sum_b h_{b,l},$$

where $s_{b,l}$ takes the value 1 if base b occurs at position l in sequence s and is 0 otherwise.

Single-template *in vitro* transcription assays (Figures 5, S4)

10 nM of linear template was mixed with 50 nM RNAP holoenzyme in transcription buffer and incubated at 37°C for 15 min to form open complexes. 1000 µM ATP and increasing concentrations of UpA (0, 10, 40, 160, and 640 µM) were added along with 10 µM of non-radiolabeled UTP plus 6 mCi of [α^{32}P]-UTP (PerkinElmer). Upon addition of nucleotides, reactions were incubated at 37°C for 10 min to allow for product formation. Reactions were stopped by addition of an equal volume of gel loading buffer (95% formamide; 25 mM EDTA; 0.025% SDS, 0.025% xylene cyanol; 0.025% bromophenol blue).

Samples were run on 20% TBE-Urea polyacrylamide gels. Bands were quantified using ImageQuant software. Observed values of UpApU / (pppApU + UpApU) were plotted vs. [UpA] / [ATP] on semi-log plot (Sigmaplot). Non-linear regression was used to fit the data to the equation: y = (ax) / (b+x); where y is UpApU / (pppApU + UpApU), x is [UpA] / [ATP], and a and b are regression parameters. The resulting fit yields the value of x for which y = 0.5. The relative efficiency $(k_{cat}/K_M)_{UpA}$ / $(k_{cat}/K_M)_{ATP}$ is equal to 1/x.

Analysis of primer-dependent initiation from the *E. coli bhsA* promoter (Figure 6B)

Culture growth and cell harvesting

Plasmids pBEN516, pKS494 or pKS497 were introduced into *E. coli* MG1655 cells. Plasmid-containing cells were grown in 25 ml of LB containing chloramphenicol (25 µg/ml) in a 125 ml

DeLong flask (Bellco Glass) at 37°C and harvested 5, 9, 14, or 21 h after cells had entered stationary phase (OD600 of ~3.3, ~3.1, ~2.9, and ~2.6, respectively). Cell suspensions were removed to 2 ml microcentrifuge tubes (Axygen), cells were collected by centrifugation (15,000 rpm; 30 s; 4°C), and cell pellets were stored at -80°C.

RNA isolation

Cells were resuspended in 0.6 ml of TRI-Reagent, incubated at 70°C for 10 min, and the cell lysate was centrifuged to remove insoluble material (10 min; 21,000 x *g*; 4°C). The supernatant was transferred to a fresh tube, ethanol was added to a final concentration of 60.5%, and the mixture was applied to a Direct-zol spin column. DNase I treatment was performed on-column according to the manufacturer's recommendations. RNA was eluted from the column using nuclease-free water that had been heated to 70°C (3 x 30 µl elutions; total volume of eluate = 90 µl). RNA was treated with 2 U TURBO DNase at 37°C for 1 h to remove residual DNA. Samples were extracted with acid phenol:chloroform, RNA was recovered by ethanol precipitation and resuspended in RNase-free water.

Primer extension analysis

Assays were performed essentially as described in (10). 10 µg of RNA was combined with 3 µM of primer k711 (5'-radiolabeled using PNK and [γ^{32}P]-ATP). The RNA-primer mixture was heated to 95°C for 10 min, slowly cooled to 40°C (0.1°C/s), incubated at 40°C for 10 min, and cooled to 4°C using a thermal cycler (Biorad). Next, 10 U of AMV reverse transcriptase (NEB) was added, reactions were incubated at 55°C for 60 min, heated to 90°C for 10 min, cooled to 4°C for 30 min, and mixed with 10 µl of 2x RNA loading buffer (95% formamide; 0.5 mM EDTA, pH 8; 0.025% SDS; 0.0025% bromophenol blue; and 0.0025% xylene cyanol). Nucleic acids were separated by electrophoresis on 8%, 7M urea slab gels (equilibrated and run in 1x TBE) and radiolabeled products were visualized by storage phosphor imaging. Band assignments were made by comparison to a DNA sequence ladder prepared using primer k711 and pBEN516 as template (Affymetrix Sequenase DNA sequencing kit, version 2).

Structure determination (Figures 7, S7, Table S2)

The nucleic-acid scaffold for assembly of *Thermus thermophilus* RPo was prepared from synthetic oligonucleotides (Sangon Biotech) by an annealing procedure (95°C, 5 min followed by 2°C-step cooling to 25°C) in 5 mM Tris-HCl (pH 8.0), 200 mM NaCl, and 10 mM $MgCl_2$ (nontemplate strand for all structures: 5'-TATAATGGGAGCTGTCACGGATGCAGG-3'; template strand for RPo[$A_{TSS-2}A_{TSS-1}T_{TSS}$]-UpA-CMPcPP: 5'-CCTGCATCCGTGAGTAAAG-3'; template strand for

RPo[$A_{TSS-2}C_{TSS-1}C_{TSS}$]-GpG-CMPcPP: 5'-CCTGCATCCGTGAGCCAAG-3'; template strand for RPo[$T_{TSS-2}C_{TSS-1}C_{TSS}$]-GpG-CMPcPP: 5'-CCTGCATCCGTGAGCCTAG-3').

For each structure, *T. thermophilus* RPo was reconstituted by mixing *T. thermophilus* RNAP holoenzyme purified as described in (18) and the nucleic-acid scaffold at a 1:1.2 molar ratio and incubating for 1h at 22°C. Crystals of *T. thermophilus* RPo were obtained and handled essentially as in (18). The primer (UpA or GpG) and CMPcPP were subsequently soaked into RPo crystals by addition of 0.2 µl 100 mM primer (UpA or GpG) and 0.2 µl 50 mM CMPcPP in RNase-free water to the crystallization drops (2 µl) and incubation for 30 min at 22ºC. Crystals were transferred in a stepwise fashion to reservoir solution (0.2 M KCl; 0.05 M $MgCl_2$; 0.1 M Tris-HCl, pH 7.9; 9% PEG 4000) containing 0.5%, 1%, 5%, 10%, and 17.5% (v/v) (2R, 3R)-(-)-2,3-butanediol, and cooled in liquid nitrogen.

Diffraction data were collected at Shanghai Synchrotron Radiation Facility (SSRF) beamlines 17U and processed using HKL2000 (36). The structures were solved by molecular replacement with Phaser MR in Phenix using one molecule of RNAP holoenzyme from the structure of *T. thermophilus* RPo (PDB:4G7H) as the search model (18, 37). Early-stage rigid-body refinement of the RNAP molecule revealed good electron density signals for the primer (UpA or GpG) and CMPcPP. Cycles of iterative model building with Coot and refinement with Phenix were performed (38, 39). The models of the primer (UpA or GpG) and CMPcPP were built into the map at later stage of refinement.

The final model of RPo[$A_{TSS-2}A_{TSS-1}T_{TSS}$]-UpA-CMPcPP was refined to R_{work} and R_{free} of 0.205 and 0.245, respectively. The final model of RPo[$A_{TSS-2}C_{TSS-1}C_{TSS}$]-GpG-CMPcPP was refined to R_{work} and R_{free} of 0.187 and 0.252, respectively. The final model of RPo[$T_{TSS-2}C_{TSS-1}C_{TSS}$]-GpG was refined to R_{work} and R_{free} of 0.194 and 0.246, respectively.

Quantification and statistical analysis

The number of replicates and statistical test procedures are in the figure legends.

Data and software availability

Sequencing reads have been deposited in the NIH/NCBI Sequence Read Archive under the study accession number PRJNA718578. Logos were generated using Logomaker (35) via custom Python scripts. Source code and documentation are provided at http://www.github.com/jbkinney/20_nickels. Structures of RPo[$A_{TSS-2}A_{TSS-1}T_{TSS}$]-UpA-CMPcPP, RPo[$A_{TSS-2}C_{TSS-1}C_{TSS}$]-GpG-CMPcPP, and RPo[$T_{TSS-2}C_{TSS-1}C_{TSS}$]-GpG have been deposited in the Protein Data Bank (PBD) under identification codes 7EH0, 7EH1, and 7EH2, respectively.

Acknowledgements

Work was supported by National Natural Science Foundation of China grant 31822001 (YZ) and National Institutes of Health grants GM124976 (PS), GM133777 (JBK), GM041376 (RHE), and GM118059 (BEN).

References

1. D. J. Hoffman, S. K. Niyogi, RNA initiation with dinucleoside monophosphates during transcription of bacteriophage T4 DNA with RNA polymerase of *Escherichia coli*. *Proc Natl Acad Sci U S A* **70**, 574-578 (1973).
2. E. G. Minkley, D. Pribnow, Transcription of the early region of bacteriophage T7: selective initiation with dinucleotides. *J Mol Biol* **77**, 255-277 (1973).
3. J. P. Dausse, A. Sentenac, P. Fromageot, Interaction of RNA Polymerase from *Escherichia coli* with DNA: Analysis of T7 DNA Early-Promoter Sites. *Eur J Biochem* **57**, 569-578 (1975).
4. P. P. Di Nocera, A. Avitabile, F. Blasi, In vitro transcription of the *Escherichia coli* histidine operon primed by dinucleotides. Effect of the first histidine biosynthetic enzyme. *J Biol Chem* **250**, 8376-8381 (1975).
5. W. J. Smagowicz, K. H. Scheit, Primed abortive initiation of RNA synthesis by *E. coli* RNA polymerase on T7 DNA. Steady state kinetic studies. *Nucleic Acids Res* **5**, 1919-1932 (1978).
6. W. Smagowicz, K. Scheit, The properties of ATP-analogs in initiation of RNA synthesis catalyzed by RNA polymerase from *E. coli*. *Nucleic Acids Res* **9**, 2397-2410 (1981).
7. N. Ruetsch, D. Dennis, RNA polymerase. Limit cognate primer for initiation and stable ternary complex formation. *J Biol Chem* **262**, 1674-1679 (1987).
8. S. R. Goldman *et al.*, NanoRNAs prime transcription initiation *in vivo*. *Mol Cell* **42**, 817-825 (2011).
9. I. O. Vvedenskaya *et al.*, Growth phase-dependent control of transcription start site selection and gene expression by nanoRNAs. *Genes Dev* **26**, 1498-1507 (2012).
10. S. Y. Druzhinin *et al.*, A Conserved Pattern of Primer-Dependent Transcription Initiation in *Escherichia coli* and *Vibrio cholerae* Revealed by 5' RNA-seq. *PLoS Genet* **11**, e1005348 (2015).
11. B. E. Nickels, A new way to start: nanoRNA-mediated priming of transcription initiation. *Transcription* **3**, 300-304 (2012).
12. I. O. Vvedenskaya *et al.*, Massively Systematic Transcript End Readout, "MASTER": Transcription Start Site Selection, Transcriptional Slippage, and Transcript Yields. *Mol Cell* **60**, 953-965 (2015).
13. I. O. Vvedenskaya, S. R. Goldman, B. E. Nickels, Analysis of Bacterial Transcription by "Massively Systematic Transcript End Readout," MASTER. *Methods Enzymol* **612**, 269-302 (2018).
14. J. T. Winkelman *et al.*, Multiplexed protein-DNA cross-linking: Scrunching in transcription start site selection. *Science* **351**, 1090-1093 (2016).
15. I. O. Vvedenskaya *et al.*, Interactions between RNA polymerase and the core recognition element are a determinant of transcription start site selection. *Proc Natl Acad Sci U S A* **113**, E2899-2905 (2016).
16. I. O. Vvedenskaya *et al.*, CapZyme-Seq Comprehensively Defines Promoter-Sequence Determinants for RNA 5' Capping with NAD^+. *Mol Cell* **70**, 553-564 e559 (2018).
17. E. Nudler, E. Avetissova, N. Korzheva, A. Mustaev, Characterization of protein-nucleic acid interactions that are required for transcription processivity. *Methods Enzymol* **371**, 179-190 (2003).
18. Y. Zhang *et al.*, Structural basis of transcription initiation. *Science* **338**, 1076-1080 (2012).
19. M. L. Gleghorn, E. K. Davydova, R. Basu, L. B. Rothman-Denes, K. S. Murakami, X-ray crystal structures elucidate the nucleotidyl transfer reaction of transcript initiation using two nucleotides. *Proc Natl Acad Sci U S A* **108**, 3566-3571 (2011).
20. R. S. Basu *et al.*, Structural basis of transcription initiation by bacterial RNA polymerase holoenzyme. *J Biol Chem* **289**, 24549-24559 (2014).
21. S. Shuman, RNA capping: progress and prospects. *RNA* **21**, 735-737 (2015).

22. A. Jaschke, K. Hofer, G. Nubel, J. Frindert, Cap-like structures in bacterial RNA and epitranscriptomic modification. *Curr Opin Microbiol* **30**, 44-49 (2016).
23. A. Ramanathan, G. B. Robb, S. H. Chan, mRNA capping: biological functions and applications. *Nucleic Acids Res* **44**, 7511-7526 (2016).
24. K. Hofer, A. Jaschke, Epitranscriptomics: RNA Modifications in Bacteria and Archaea. *Microbiol Spectr* **6** (2018).
25. J. G. Bird *et al.*, The mechanism of RNA 5' capping with NAD^+, NADH and desphospho-CoA. *Nature* **535**, 444-447 (2016).
26. I. Barvik, D. Rejman, N. Panova, H. Sanderova, L. Krasny, Non-canonical transcription initiation: the expanding universe of transcription initiating substrates. *FEMS Microbiol Rev* **41**, 131-138 (2017).
27. C. Julius, Y. Yuzenkova, Bacterial RNA polymerase caps RNA with various cofactors and cell wall precursors. *Nucleic Acids Res* **45**, 8282-8290 (2017).
28. J. G. Bird *et al.*, Highly efficient 5' capping of mitochondrial RNA with NAD^+ and NADH by yeast and human mitochondrial RNA polymerase. *Elife* **7** (2018).
29. D. J. Luciano, R. Levenson-Palmer, J. G. Belasco, Stresses that Raise Np4A Levels Induce Protective Nucleoside Tetraphosphate Capping of Bacterial RNA. *Mol Cell* **75**, 957-966 e958 (2019).
30. D. J. Luciano, J. G. Belasco, Np4A alarmones function in bacteria as precursors to RNA caps. *Proc Natl Acad Sci U S A* **117**, 3560-3567 (2020).
31. V. Svetlov, I. Artsimovitch, Purification of bacterial RNA polymerase: tools and protocols. *Methods Mol Biol* **1276**, 13-29 (2015).
32. I. Artsimovitch, V. Svetlov, K. S. Murakami, R. Landick, Co-overexpression of Escherichia coli RNA polymerase subunits allows isolation and analysis of mutant enzymes lacking lineage-specific sequence insertions. *J Biol Chem* **278**, 12344-12355 (2003).
33. M. T. Marr, J. W. Roberts, Promoter recognition as measured by binding of polymerase to nontemplate strand oligonucleotide. *Science* **276**, 1258-1260 (1997).
34. J. T. Winkelman, P. Chandrangsu, W. Ross, R. L. Gourse, Open complex scrunching before nucleotide addition accounts for the unusual transcription start site of *E. coli* ribosomal RNA promoters. *Proc Natl Acad Sci U S A* **113**, E1787-E1795 (2016).
35. A. Tareen, J. B. Kinney, Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272-2274 (2020).
36. Z. Otwinowski, W. Minor, Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* **276**, 307-326 (1997).
37. A. J. McCoy *et al.*, Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007).
38. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126-2132 (2004).
39. P. D. Adams *et al.*, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221 (2010).

Figures

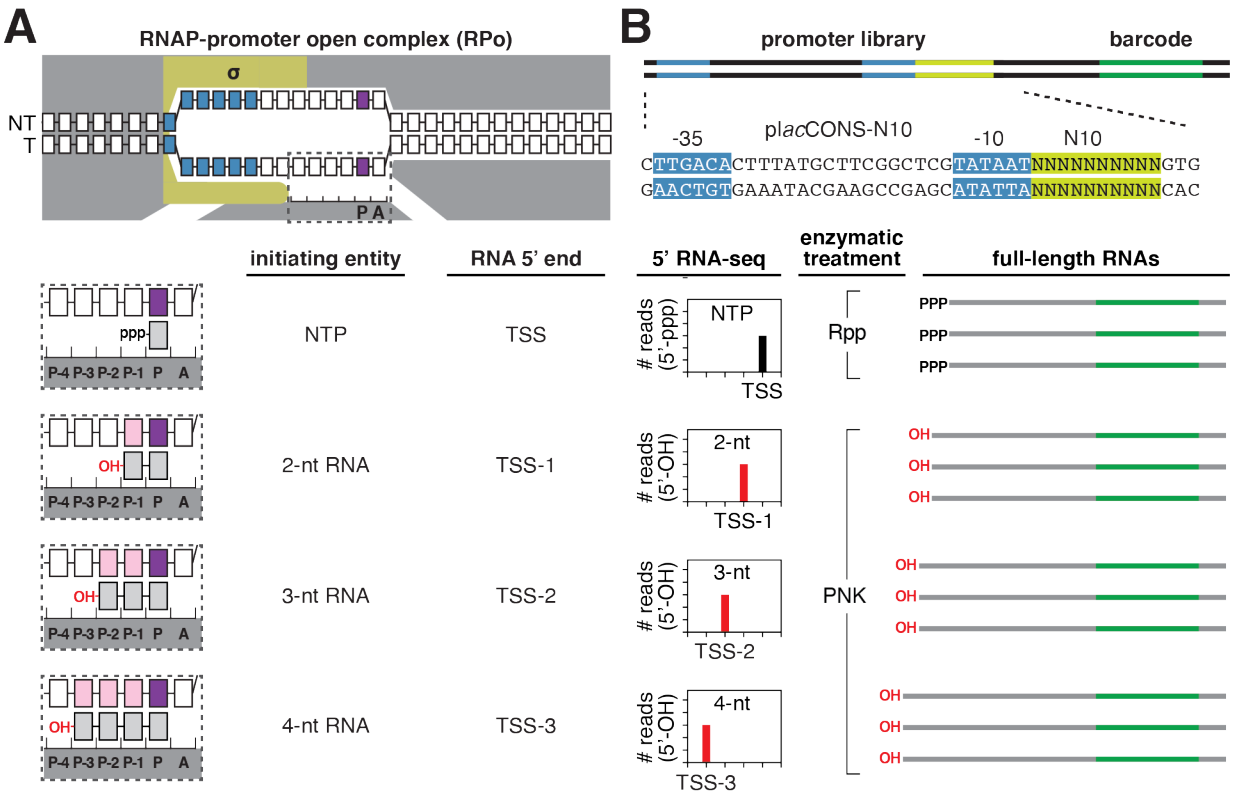


Figure 1. Use of massively systematic transcript end readout, "MASTER," to monitor primer-independent and primer-dependent transcription initiation.

A. Binding of initiating entities to DNA template-strand nucleotides in primer-independent and primer-dependent transcription initiation. Top: RNAP-promoter open complex (RPo). Bottom: Enlarged view of initiating entities bound to template-strand nucleotides in the RNAP active center. Dark gray, RNAP; yellow, σ; blue, -10-element nucleotides; purple, transcription start site (TSS) nucleotides; light gray, RNA nucleotides; pink, primer-binding nucleotides at positions TSS-1, TSS-2 or TSS-3; white boxes, DNA nucleotides; NT, nontemplate-strand nucleotides; T, template-strand nucleotides. P-3, P-2, P-1, and P, RNAP active-center initiating entity binding sites; A, RNAP active-center extending NTP binding site. Unwound transcription bubble in RPo indicated by raised and lowered nucleotides.

B. Analysis of primer-independent and primer-dependent initiation using MASTER. Top: DNA fragment containing MASTER template library. Light green, randomized nucleotides in the promoter region; dark green, transcribed-region barcode. Bottom: 5′ RNA-seq analysis of RNA products generated from the promoter library by primer-independent, NTP-dependent initiation (Rpp treatment) and primer-dependent initiaion (PNK treatment).

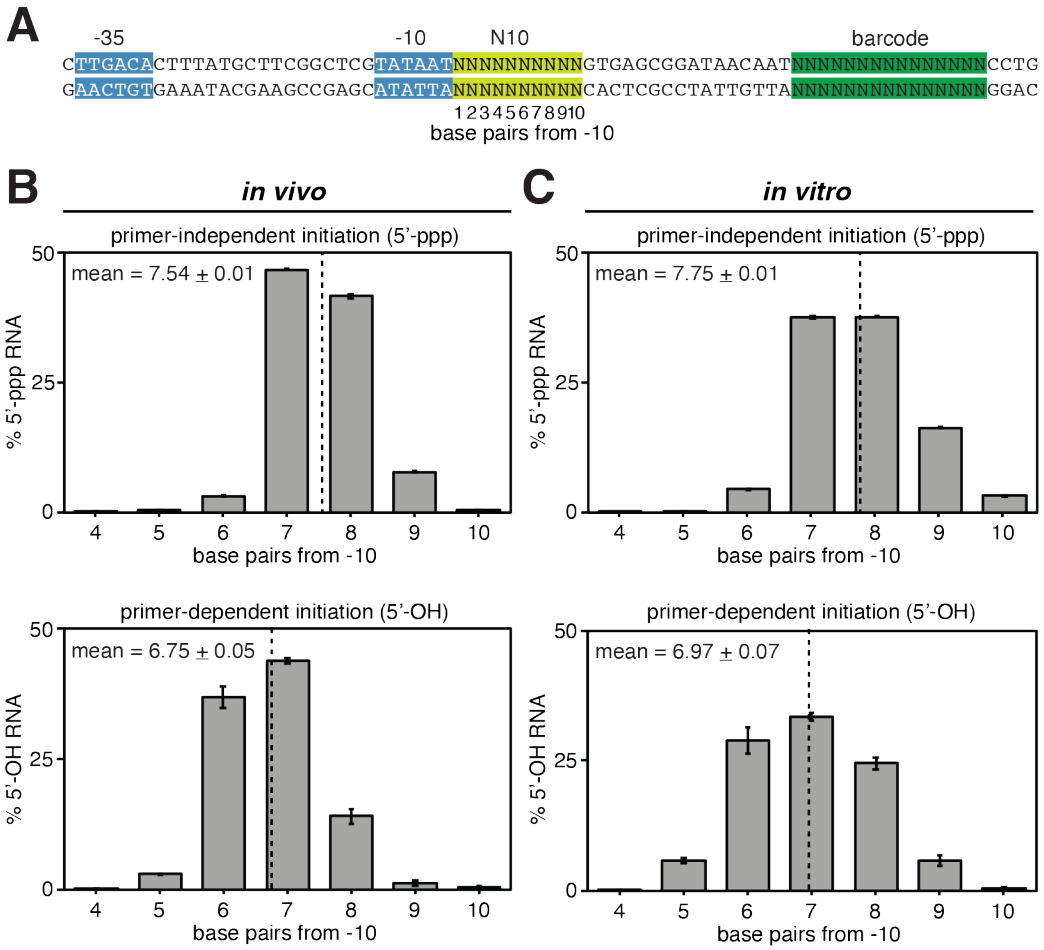


Figure 2. Distributions of 5′-end sequences for RNAs generated in primer-independent initiation and primer-dependent initiation *in vivo* and *in vitro*.

A. *lac*CONS-N10 library. Base pairs in the N10 region are numbered based on their position relative to the promoter -10 element. Colors as in Figure 1B.

B-C. RNA 5′-end distribution histograms (mean ± SD, N = 3) for RNAs generated by primer-independent initiation (top) or primer-dependent initiation (bottom) in stationary-phase *E. coli* cells (panel B) or *in vitro*, with the dinucleotide primer UpA (panel C). Dashed line indicates the mean 5′-end position (mean ± SD, N = 3).

A

primer	% 5′-OH RNA	primer binding site				
		5-6	6-7	7-8	8-9	9-10
NpN		3.0	37.5	44.1	13.8	1.4
ApA	2.1	0.1	0.8	0.8	0.4	0.1
ApC	2.2	0.1	0.8	1.2	0.2	0.0
ApG	5.0	0.0	1.4	2.1	1.2	0.3
ApU	1.2	0.1	0.4	0.3	0.2	0.1
CpA	2.1	0.1	0.8	1.0	0.3	0.0
CpC	2.9	0.0	1.1	1.5	0.2	0.0
CpG	1.9	0.1	0.7	0.9	0.2	0.0
CpU	2.3	0.1	1.0	1.0	0.2	0.2
GpA	6.3	0.0	2.4	2.5	1.1	0.0
GpC	5.8	0.0	1.9	2.7	1.1	0.1
GpG	**16.9**	**0.1**	**7.2**	**7.4**	**1.9**	**0.1**
GpU	5.3	0.1	1.7	2.3	1.1	0.1
UpA	**27.0**	**0.9**	**10.9**	**10.9**	**4.0**	**0.3**
UpC	7.7	0.2	2.2	4.6	0.7	0.0
UpG	5.3	0.9	1.8	2.1	0.4	0.1
UpU	6.2	0.4	2.3	2.8	0.6	0.0

B

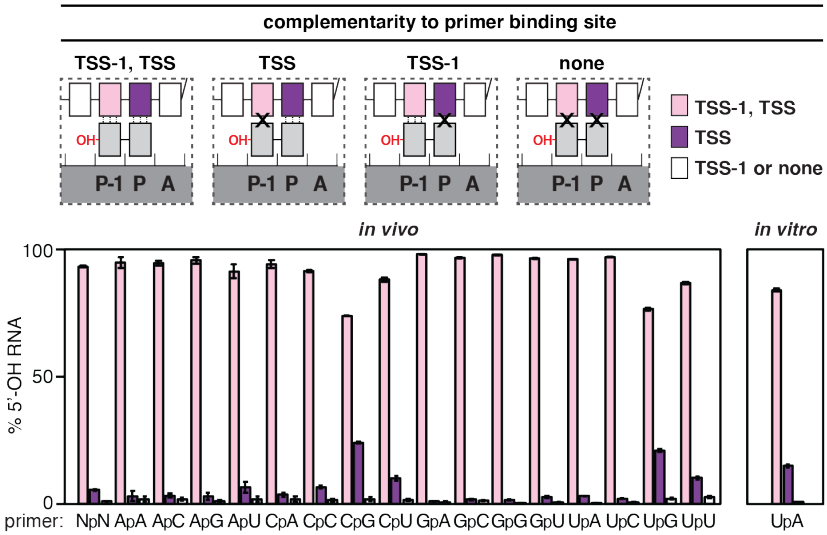


Figure 3. Promoter-sequence dependence of primer-dependent initiation: primer binding site.
A. Relative usage of dinucleotides in primer-dependent initiation in stationary-phase *E. coli* cells. Values represent the percentage of total 5′-OH RNAs generated using each of the 16 dinucleotide primers (mean, N = 3). Bold, dinucleotides preferentially used as primers.
B. Complementarity between the primer binding site and dinucleotide in primer-dependent initiation. Top: primer-dependent initiation involving template-strand complementarity to both 5′ and 3′ nucleotides of primer (TSS-1, TSS), template-strand complementarity to only 3′ nucleotide of primer (TSS), template-strand complementarity to only 5′ nucleotide of primer (TSS-1), or no template-strand complementarity to primer (none). Three vertical lines, complementarity; X, non-complementarity. Other symbols and colors as in Figure 1. Bottom: percentage of primer-dependent initiation involving complementarity to both 5′ and 3′ nucleotides of primer (TSS-1, TSS; pink), complementarity to only 3′ nucleotide of primer (TSS; purple), or template-strand complementarity to only 5′ nucleotide of primer or no template-strand complementarity to primer (TSS-1 or none; white) in stationary-phase *E. coli* cells (left) or *in vitro*, with the dinucleotide primer UpA (right) (mean ± SD, N = 3).

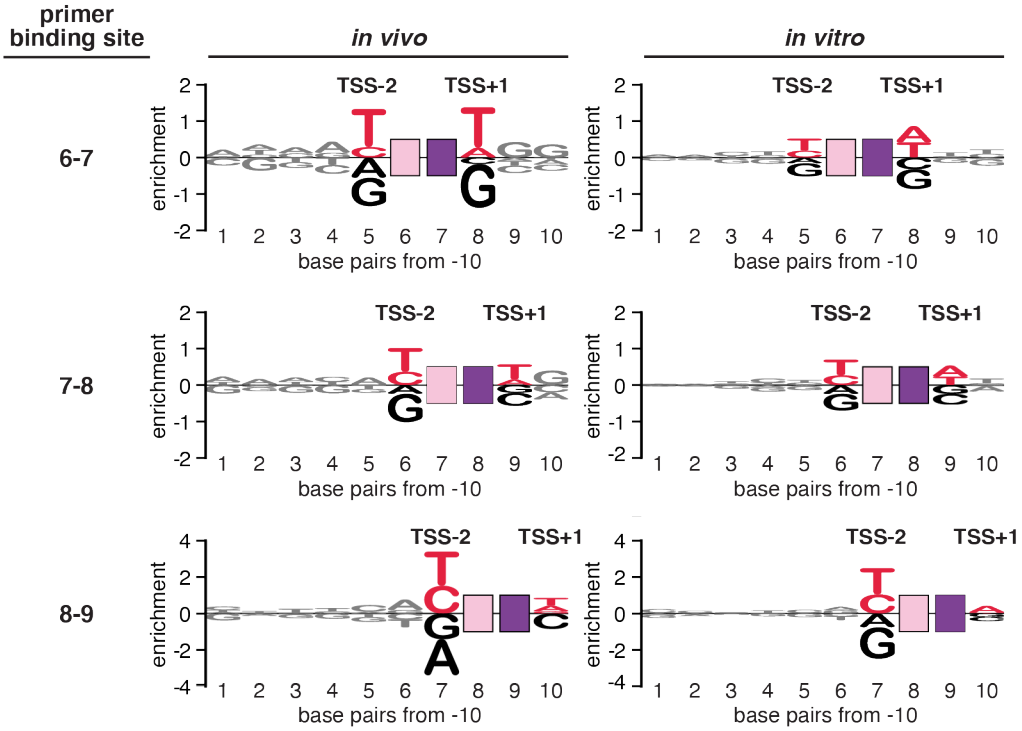


Figure 4. Promoter-sequence dependence of primer-dependent initiation: sequences flanking the primer binding site.

Sequence logo (35) for primer-dependent initiation at TSS positions 7, 8, and 9 (corresponding to primer binding sites 6-7, 7-8, and 8-9, respectively) in stationary-phase *E. coli* cells (left) or *in vitro*, with the dinucleotide primer UpA (right). The height of each base "X" at each position "Y" represents the $\log_2$ average of the % 5′-OH RNAs computed across sequences containing nontemplate-strand X at position Y. Red, consensus nucleotides; black, non-consensus nucleotides. Other symbols and colors as in Figure 1.

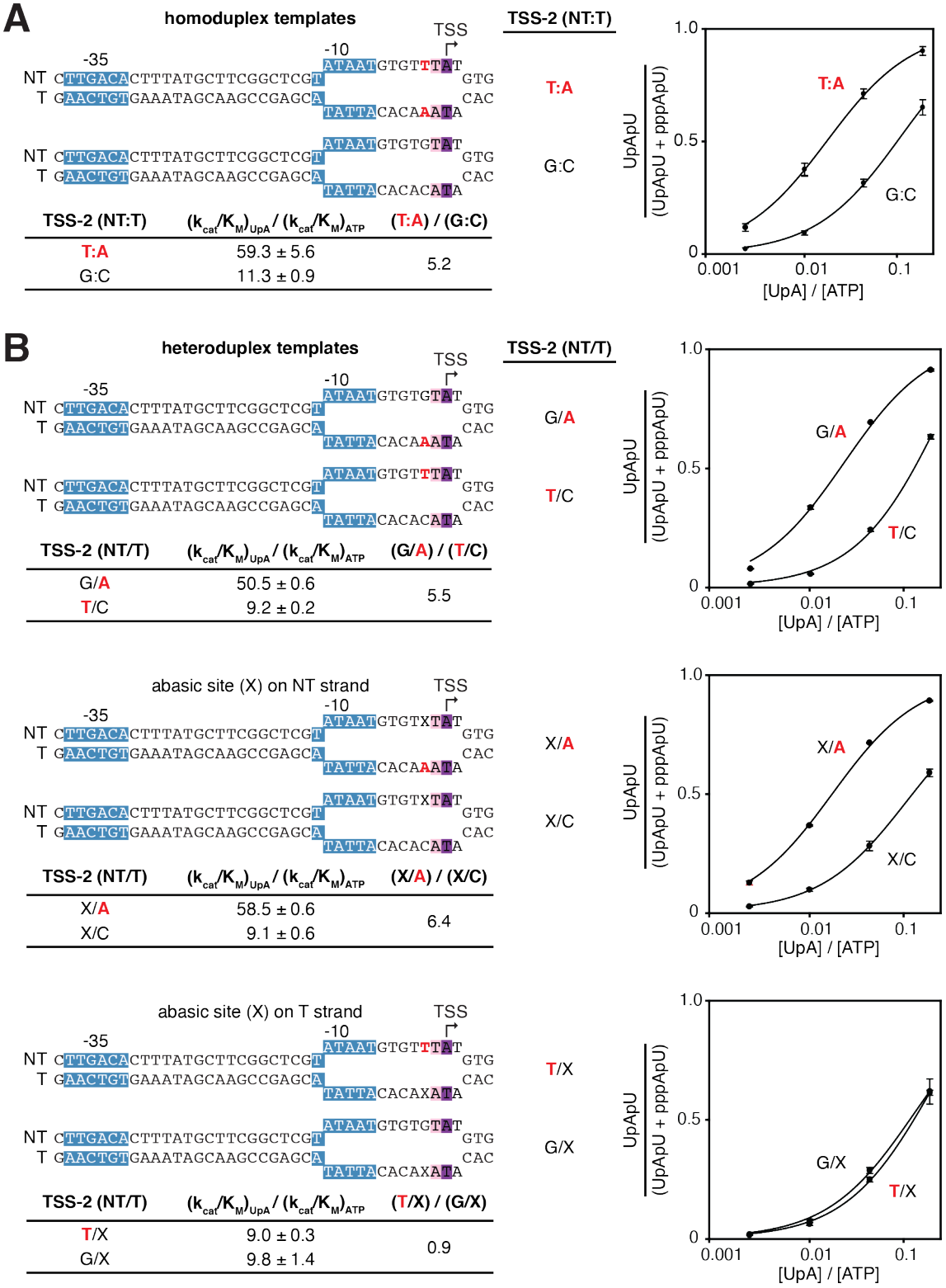


Figure 5. Promoter-sequence dependence of primer-dependent initiation *in vitro*: position TSS-2. **A.** Relative efficiencies of primer-dependent initiation vs. primer-independent initiation depends on promoter sequence at position TSS-2. Top left: homoduplex DNA templates containing consensus or

non-consensus nucleotides for primer-dependent initiation at position TSS-2. Bottom left: relative efficiencies of primer-dependent initiation with UpA vs. primer-independent initiation with ATP [$(k_{cat}/K_M)_{UpA}$ / $(k_{cat}/K_M)_{ATP}$], and ratio of $(k_{cat}/K_M)_{UpA}$ / $(k_{cat}/K_M)_{ATP}$ for the indicated templates. Right: dependence of primer-dependent initiation on [UpA] / [ATP] ratio (mean ± SD, N = 3). Red, consensus nucleotides at position TSS-2. Unwound transcription bubble in RPo indicated by raised and lowered nucleotides. Other symbols and colors as in Figure 1.

B. The template DNA strand carries sequence information at position TSS-2. Top left: DNA templates containing mismatches at position TSS-2. Templates contain a consensus nucleotide at position TSS-2 on only the nontemplate strand (T/C_{TSS-2} and T/X_{TSS-2}), only the template strand (G/A_{TSS-2} and X/A_{TSS-2}), or neither strand (X/C_{TSS-2} and G/X_{TSS-2}). Bottom left: relative efficiencies and efficiency ratios for the indicated heteroduplex templates. Right: dependence of primer-dependent initiation on [UpA] / [ATP] ratio (mean ± SD, N = 3). Red, consensus nucleotides at position TSS-2. Unwound transcription bubble in RPo indicated by raised and lowered nucleotides.

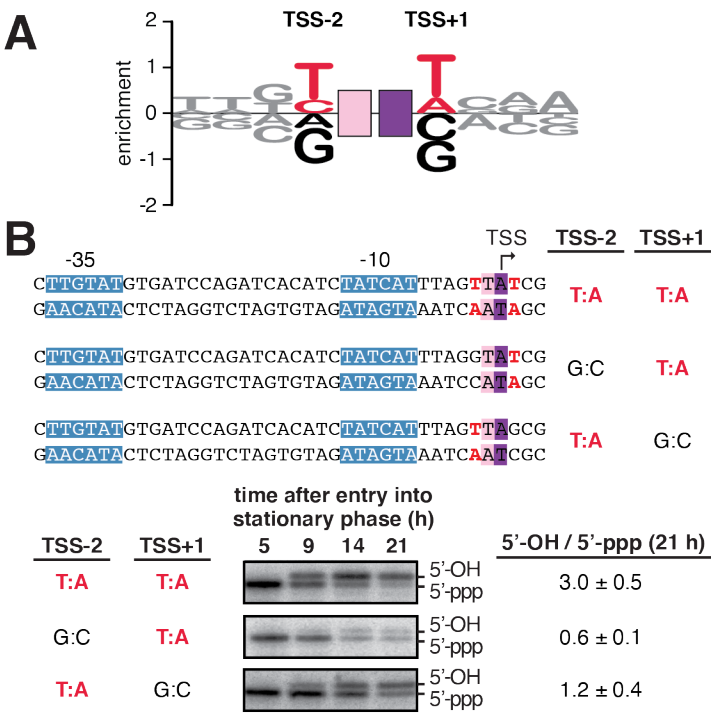


Figure 6. Promoter-sequence dependence of primer-dependent initiation: chromosomal promoters
A. Sequence logo (35) for primer-dependent initiation at TSS positions 7, 8, and 9 (corresponding to primer binding sites 6-7, 7-8, and 8-9, respectively) in stationary-phase *E. coli* cells for 93 natural, chromosomally-encoded promoters that use UpA as a primer. The height of each base "X" at each position "Y" represents the $\log_2$ average of the % 5′-OH RNAs computed across sequences containing nontemplate-strand X at position Y. Red, consensus nucleotides; black, non-consensus nucleotides. Other symbols and colors as in Figure 1.
B. Promoter-sequence dependence of primer-dependent initiation at the *E. coli bhsA* promoter. Top: sequences of DNA templates containing wild-type and mutant derivatives of *bhsA* promoter. Bottom: primer extension analysis of 5′-end lengths of *bhsA* RNAs. In primer-dependent initiation with a dinucleotide primer, the RNA product acquires one additional nucleotide at the RNA 5′ end (Figure 1). Gel shows radiolabeled cDNA products derived from primer-independent initiation (5′-ppp) and primer-dependent initiation (5′-OH) in stationary-phase *E. coli* cells. Bottom right: ratios of primer-dependent initiation vs. primer-independent initiation (mean ± SD, N = 4).

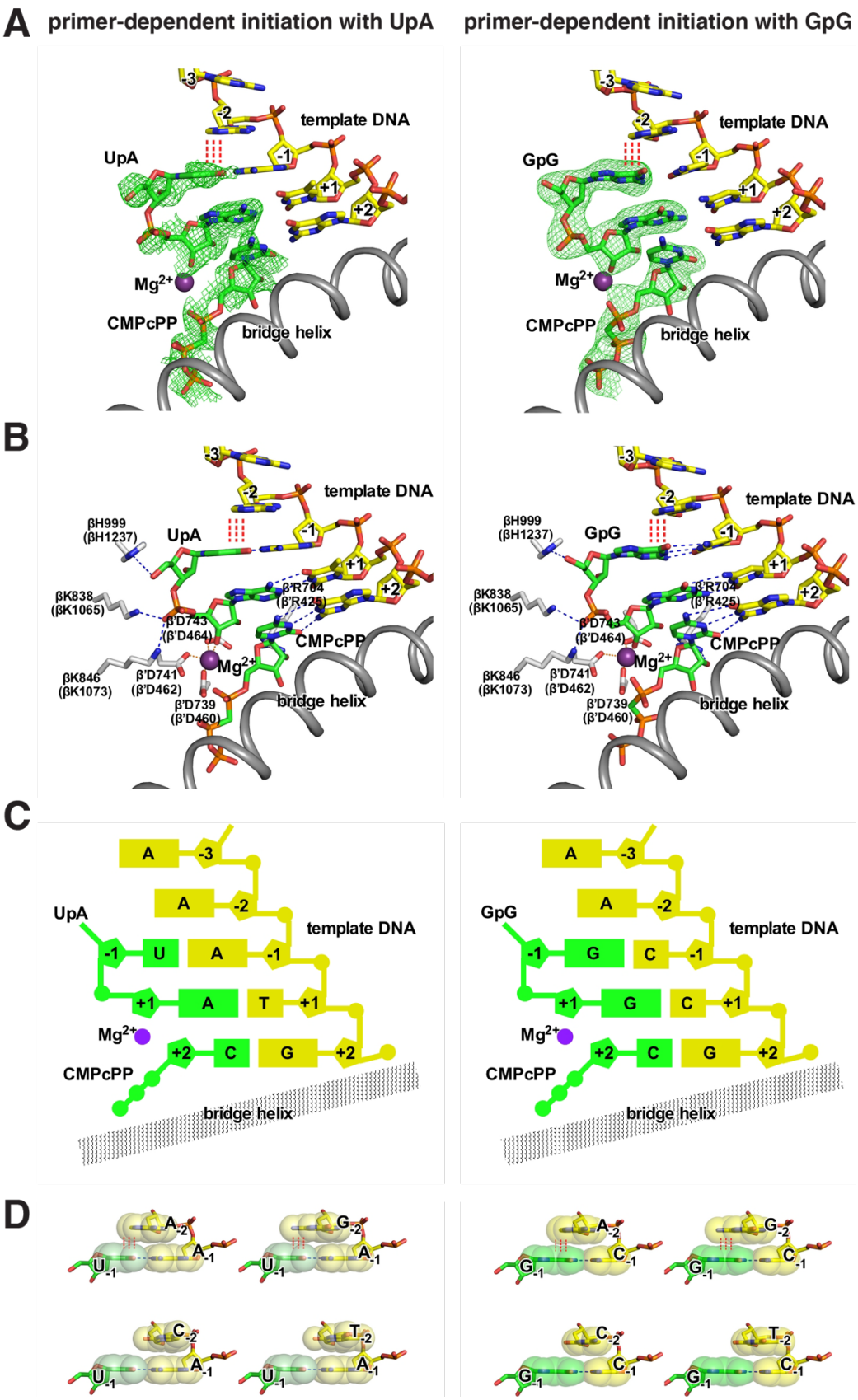


Figure 7. Structural basis of promoter-sequence dependence of primer-dependent initiation at position TSS-2.

Crystal structures of *T. thermophilus* RPo[$A_{TSS-2}A_{TSS-1}T_{TSS}$]-UpA-CMPcPP (left) and *T. thermophilus* RPo[$A_{TSS-2}C_{TSS-1}C_{TSS}$]-GpG-CMPcPP (right).

A. Experimental electron density (contoured at 2.5σ; green mesh) and atomic model for DNA template strand (yellow, red, blue, and orange for C, O, N, and P atoms), dinucleotide primer and CMPcPP (green, red, blue, and orange for C, O, N, and P atoms), RNAP active-center catalytic Mg^{2+}(I) (violet sphere), and RNAP bridge helix (gray ribbon).

B. Contacts of RNAP residues (gray, red, and blue for C, O, and N atoms) with primer and RNAP active-center catalytic Mg^{2+}(I). RNAP residues are numbered both as in *T. thermophilus* RNAP and as in *E. coli* RNAP (in parentheses).

C. Schematic summary of structures. Template-strand DNA (yellow); primer and CMPcPP (green); RNAP bridge helix (gray); RNAP active-center catalytic Mg^{2+}(I) (violet).

D. Structural basis of promoter-sequence dependence at position TSS-2. Extensive inter-chain base stacking of template-strand purine, A or G, with 5′ nucleotide of primer (upper row; red vertical dashed lines), and limited inter-chain base stacking of template-strand pyrimidine, C or T, and 5′ nucleotide of primer (lower row). The inter-chain base-stacking patterns of template-strand A with primers UpA and GpG are as observed in structures of RPo[$A_{TSS-2}A_{TSS-1}T_{TSS}$]-UpA-CMPcPP and RPo[$A_{TSS-2}C_{TSS-1}C_{TSS}$]-GpG-CMPcPP (panels A-C); the inter-chain base-stacking pattern of template-strand T with primer GpG is as observed in structure of RPo[$T_{TSS-2}C_{TSS-1}C_{TSS}$]-GpG-CMPcPP (Figure S7); the other inter-chain base-stacking patterns are modeled by analogy. Base atoms are shown as van der Waals surfaces. Colors are as in panel A.

Supplementary Figures

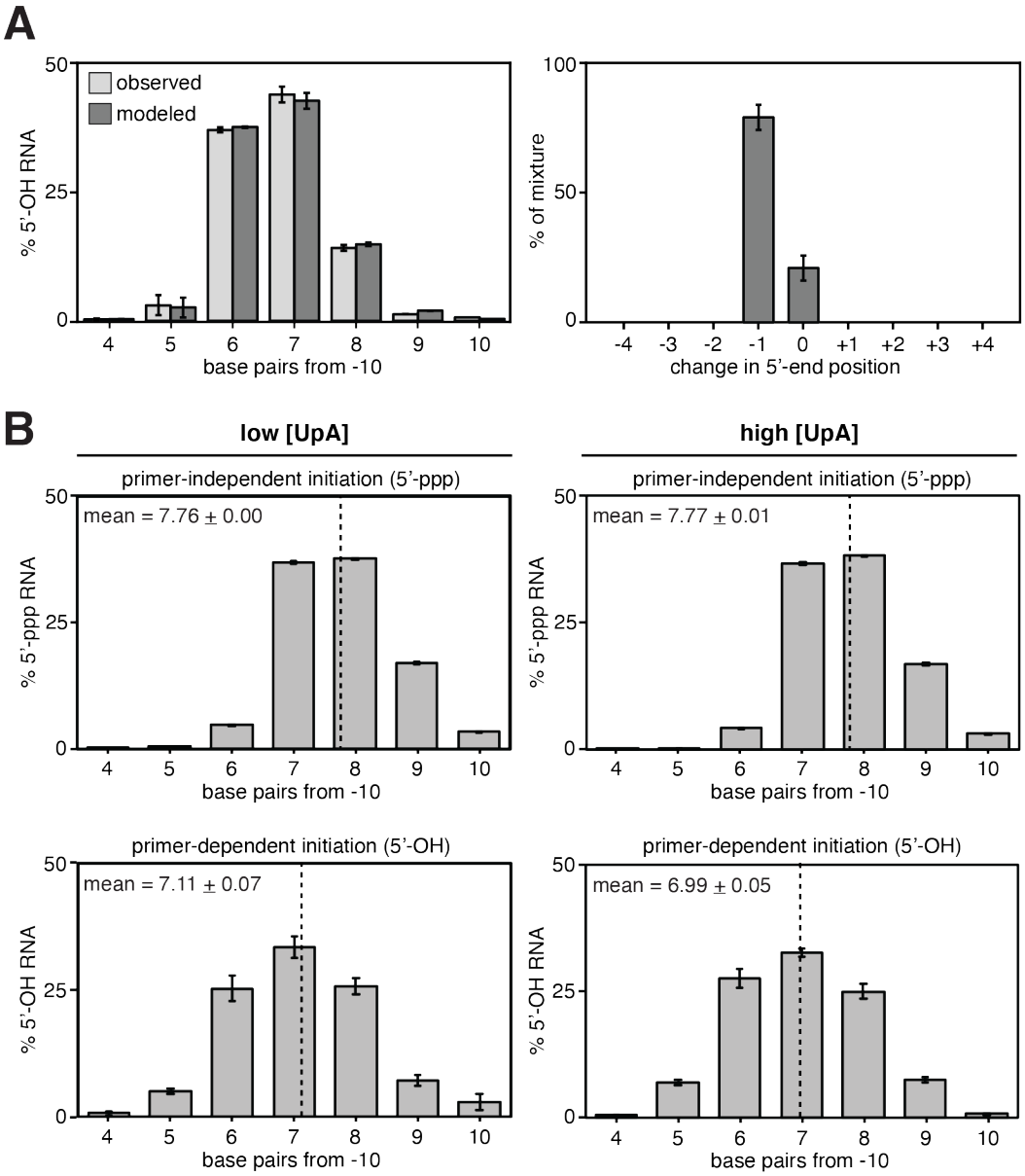


Figure S1. Distributions of 5′-end sequences for RNAs generated in primer-independent initiation and primer-dependent initiation.

A. Computational modeling of *in vivo* distributions of 5′-OH RNAs from observed *in vivo* distributions of 5′-ppp RNAs (see Figure 2B, top). Left: histograms of observed *in vivo* distributions of 5′-OH RNAs (light gray; see also Figure 2B, bottom) and modeled distributions of 5′-OH RNAs (dark gray). The modeled distributions of 5′-OH RNAs was generated by modeling distributions of 5′-OH RNAs as a mixture of 5′-ppp RNAs with positions changed by up to 4 bp upstream (-1 to -4) or downstream (+1 to +4). Right: Mixture coefficient histogram. Coefficients were inferred using least-squares regression under positivity and normalization constraints.

B. RNA 5′-end distribution histograms (mean ± SD, N = 3) for RNAs generated by primer-independent initiation or primer-dependent initiation *in vitro* (top and bottom, repectively). low UpA, [UpA] / [NTPs] = 0.04; high UpA, [UpA] / [NTPs] = 0.64. Dashed line indicates the mean 5′-end position (mean ± SD, N = 3).

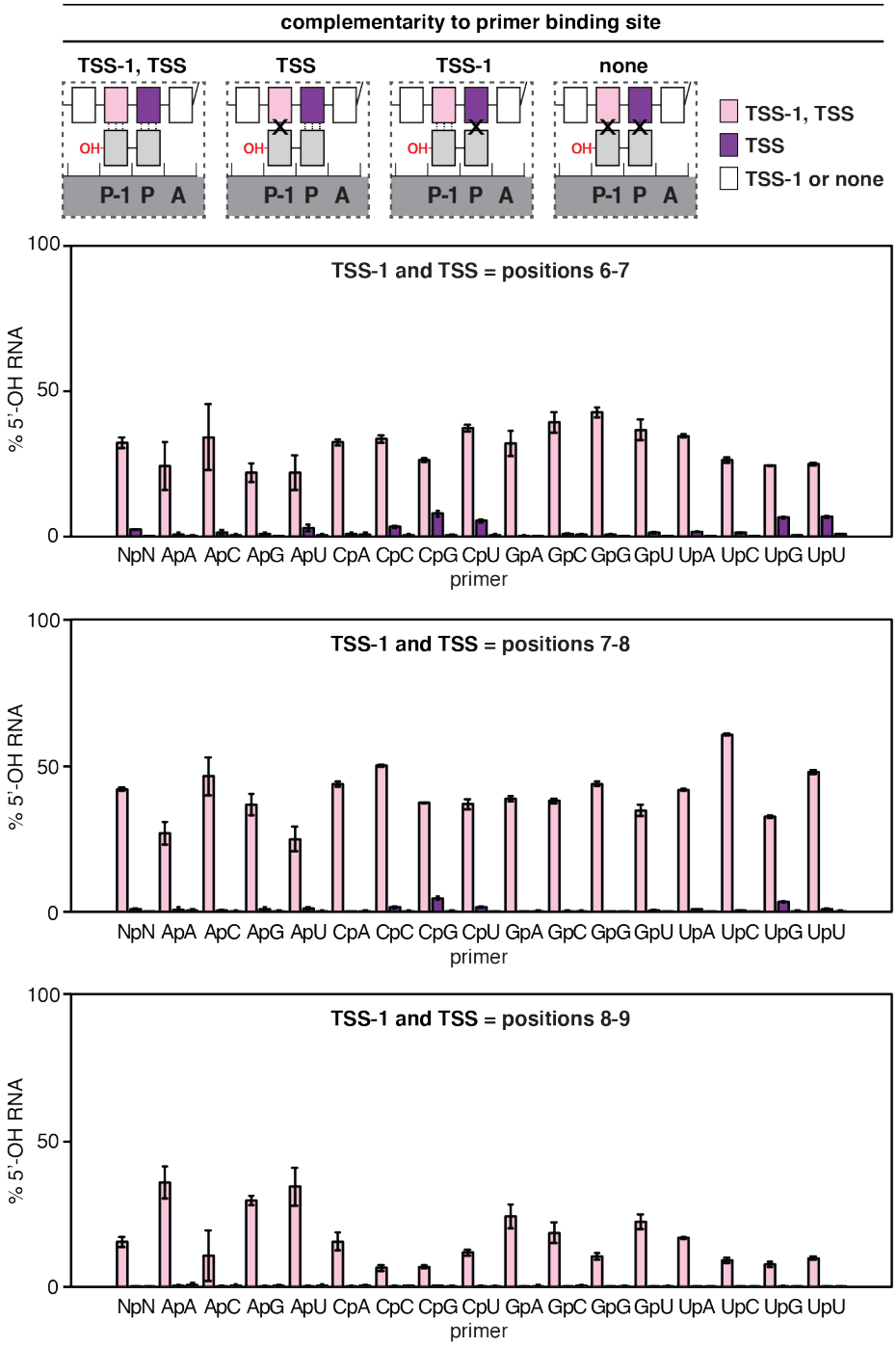


Figure S2. Promoter-sequence dependence of primer-dependent initiation in stationary-phase *E. coli* cells: primer binding site

Top: primer-dependent initiation involving template-strand complementarity to both 5′ and 3′ nucleotides of primer (TSS-1, TSS), template-strand complementarity to only 3′ nucleotide of primer (TSS), template-strand complementarity to only 5′ nucleotide of primer (TSS-1), or no template-strand complementarity to primer (none). Three vertical lines, complementarity; X, non-complementarity. Other symbols and colors as in Figure 1. Bottom: percentage of primer-dependent initiation involving complementarity to both 5′ and 3′ nucleotides of primer (TSS-1, TSS; pink), complementarity to only 3′

nucleotide of primer (TSS; purple), or template-strand complementarity to only 5′ nucleotide of primer or no template-strand complementarity to primer (TSS-1 or none; white) in stationary-phase *E. coli* cells for primer binding sites located 6-7, 7-8, or 8-9 base pairs downstream of the promoter -10 element (mean ± SD, N = 3).

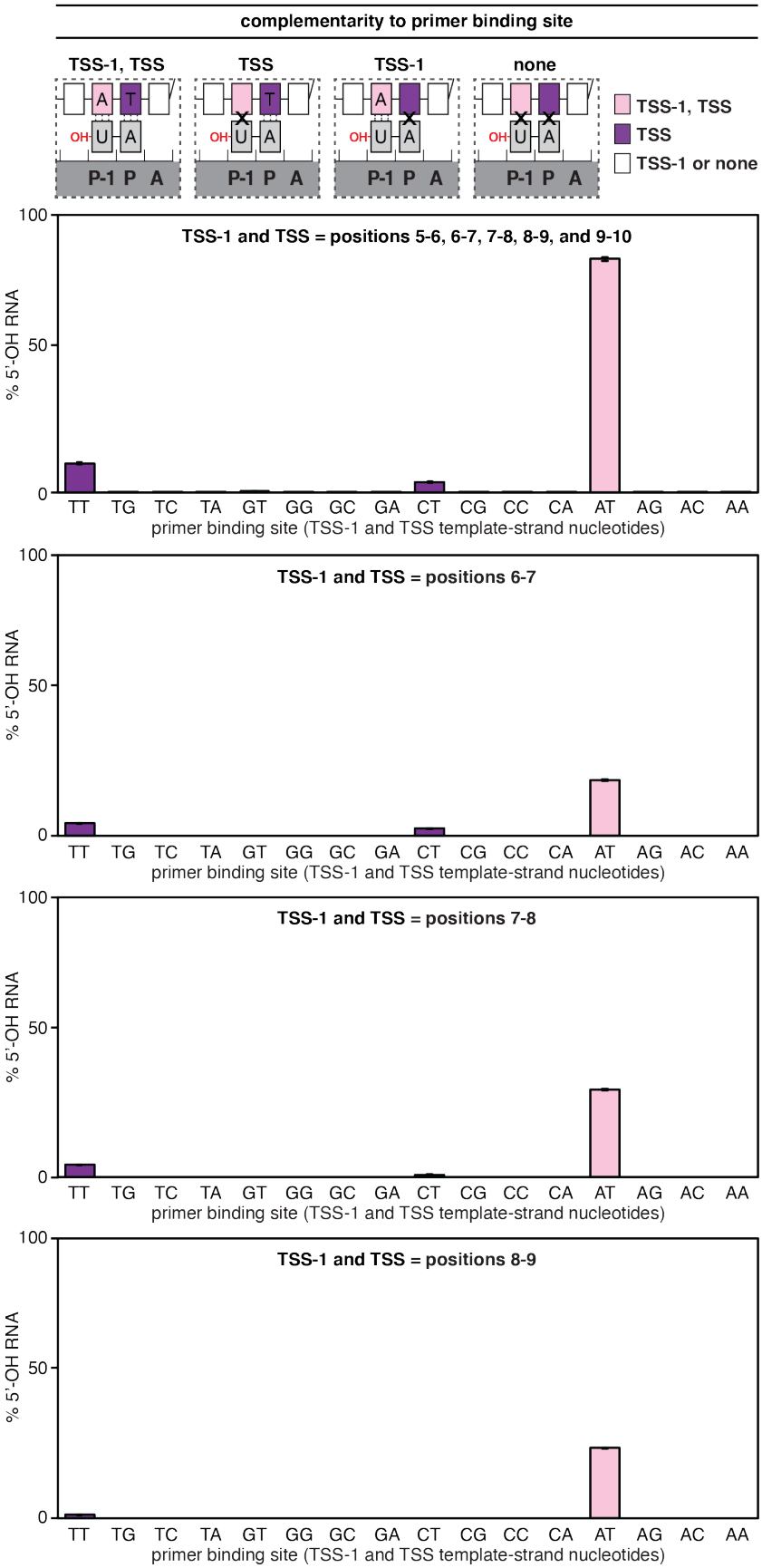


Figure S3. Promoter-sequence dependence of primer-dependent initiation *in vitro*: primer binding site

Top: primer-dependent initiation involving template-strand complementarity to both 5′ and 3′ nucleotides of UpA (TSS-1, TSS), template-strand complementarity to only 3′ nucleotide of UpA (TSS), template-strand complementarity to only 5′ nucleotide of UpA (TSS-1), or no template-strand complementarity to UpA (none). Three vertical lines, complementarity; X, non-complementarity. Other symbols and colors as in Figure 1. Bottom: percentage of primer-dependent initiation involving complementarity to both 5′ and 3′ nucleotides of UpA (TSS-1, TSS; pink), complementarity to only 3′ nucleotide of UpA (TSS; purple), or template-strand complementarity to only 5′ nucleotide of UpA or no template-strand complementarity to UpA (TSS-1 or none; white) for primer binding sites located 6-7, 7-8, or 8-9 base pairs downstream of the promoter -10 element (mean ± SD, N = 3).

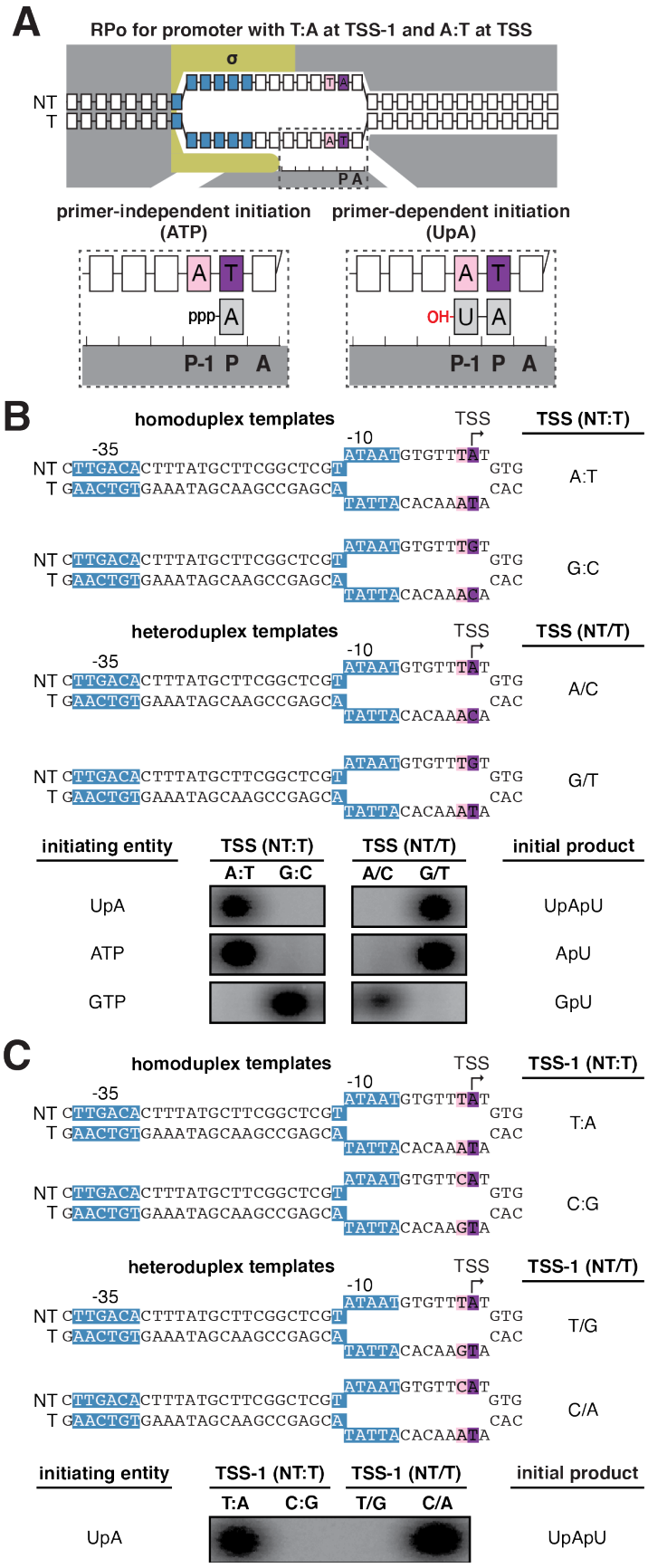


Figure S4. Promoter-sequence dependence of primer-dependent initiation *in vitro*: template strand carries sequence information at positions TSS and TSS-1

A. Binding of ATP or UpA to template-strand nucleotides in primer-independent initiation (left) and primer-dependent initiation (right). Unwound transcription bubble in RPo indicated by raised and lowered nucleotides. Other symbols and colors as in Figure 1.

B-C. Strand specificity at positions TSS and TSS-1. Top: homoduplex DNA templates. Middle: heteroduplex DNA templates containing mismatches at positions TSS (panel A) or TSS-1 (panel B). Unwound transcription bubble in RPo indicated by raised and lowered nucleotides. Bottom: radiolabeled initial RNA products generated using the indicated template in reactions containing UpA, ATP or GTP (panel B) or UpA (panel C).

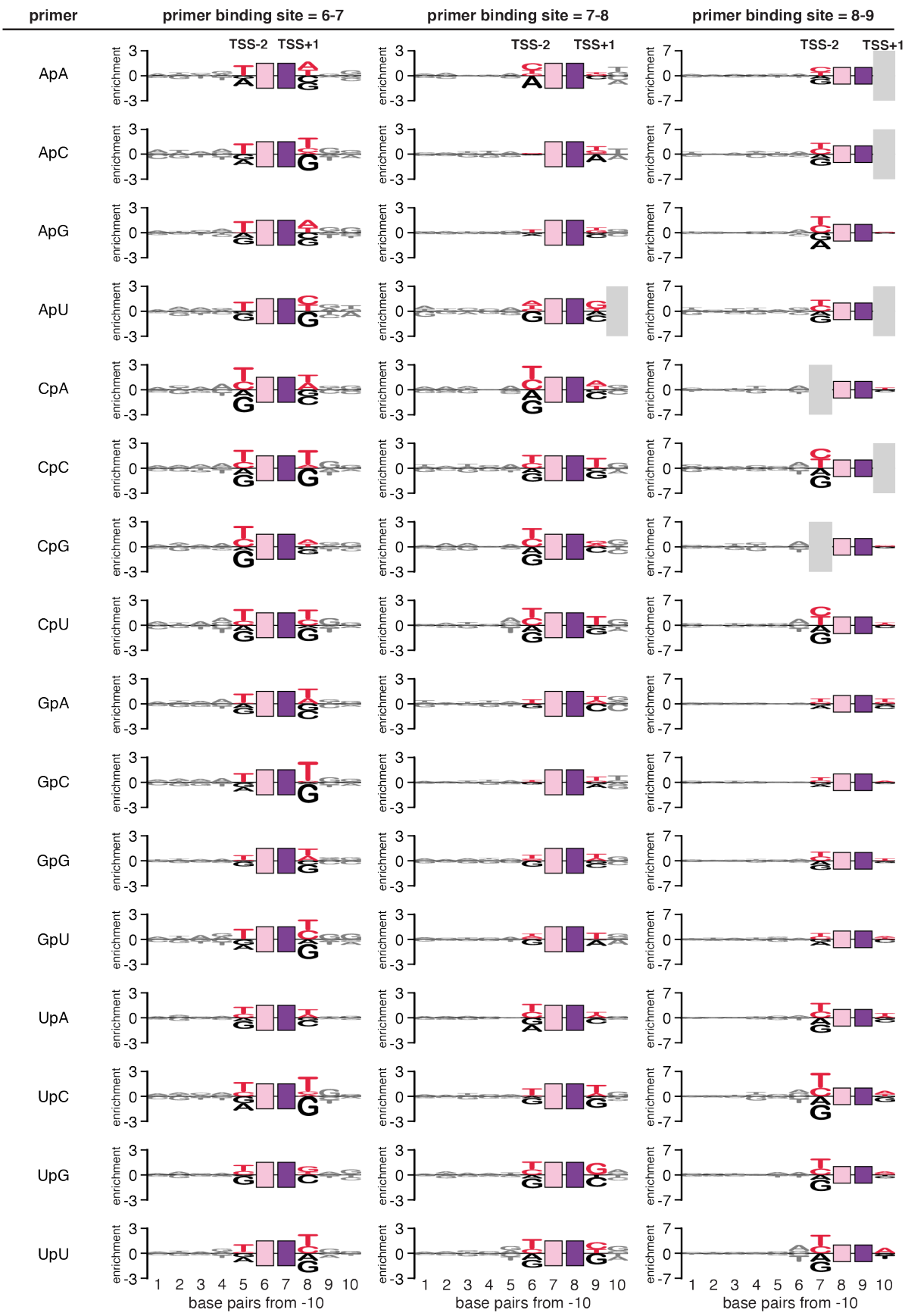


Figure S5. Promoter-sequence dependence of primer-dependent initiation *in vivo*: sequences flanking the primer binding site

Sequence logo (35) for primer-dependent initiation in stationary-phase *E. coli* cells with each of the 16 dinucleotides at TSS positions 7, 8, and 9 (corresponding to primer binding sites 6-7, 7-8, and 8-9, respectively). The height of each base "X" at each position "Y" represents the $\log_2$ average of the % 5′-OH RNAs computed across sequences containing nontemplate-strand X at position Y. Red, consensus nucleotides; black, non-consensus nucleotides. Gray box indicates positions where enrichment values could not be computed.

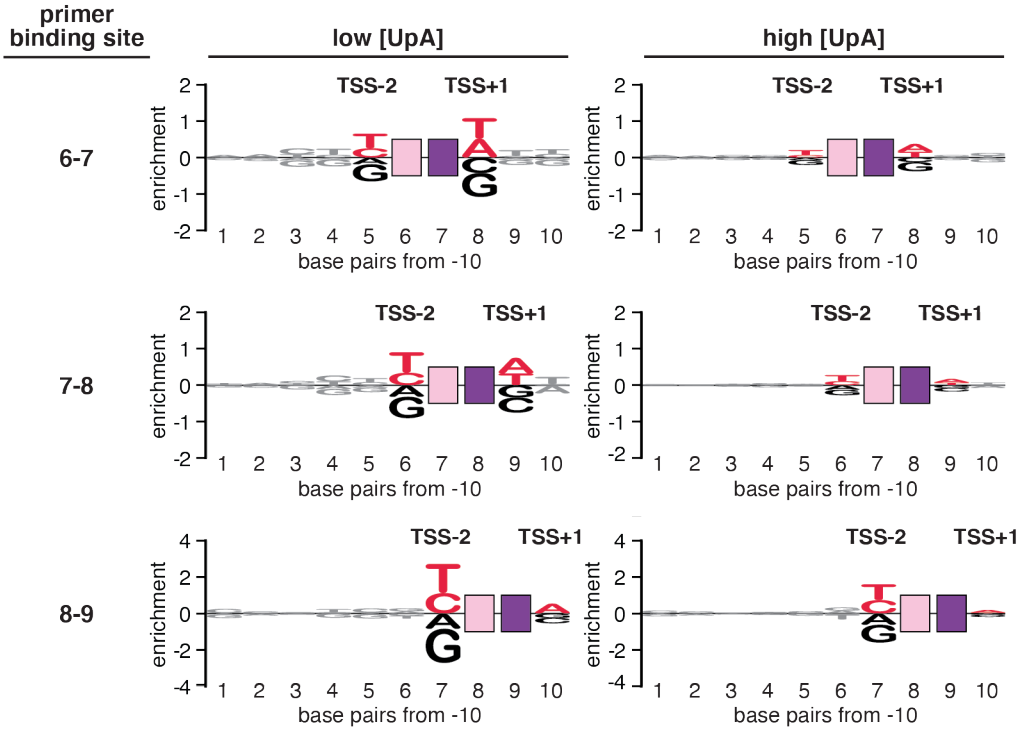


Figure S6. Promoter-sequence dependence of primer-dependent initiation *in vitro*: sequences flanking the primer binding site

Sequence logo (35) for primer-dependent initiation *in vitro* with UpA at TSS positions 7, 8, and 9 (corresponding to primer binding sites 6-7, 7-8, and 8-9, respectively). The height of each base "X" at each position "Y" represents the $\log_2$ average of the % 5′-OH RNAs computed across sequences containing nontemplate-strand X at position Y. Low UpA, [UpA] / [NTPs] = 0.04; high UpA, [UpA] / [NTPs] = 0.64. Red, consensus nucleotides; black, non-consensus nucleotides.

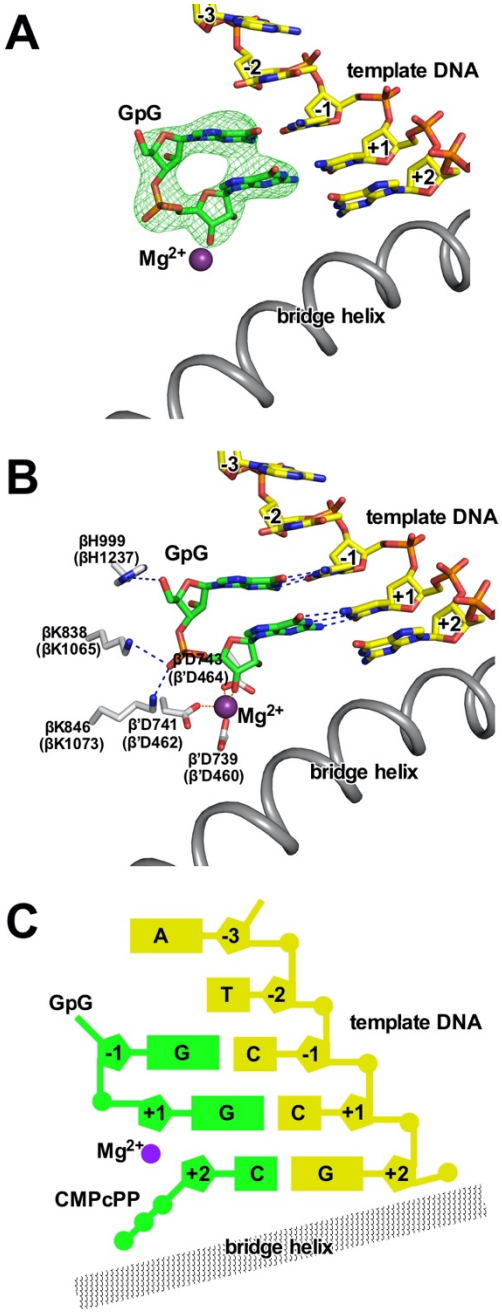


Figure S7. Structural basis of promoter-sequence dependence of primer-dependent initiation at position TSS-2

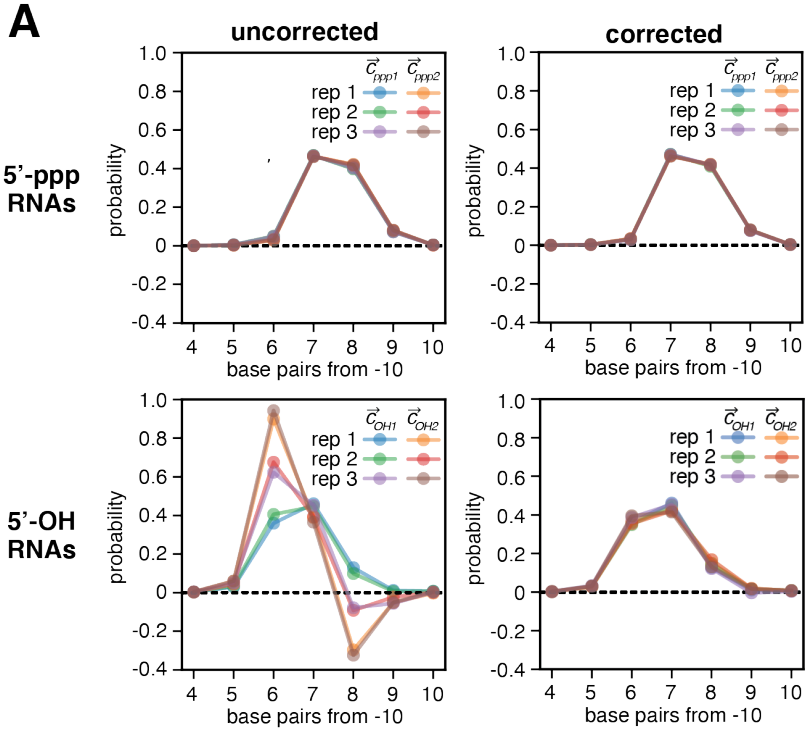
Crystal structure of *T. thermophilus* RPo[$T_{TSS-2}C_{TSS-1}C_{TSS}$]-GpG-CMPcPP

A. Experimental electron density (contoured at 2.5σ; green mesh) and atomic model for DNA template strand (yellow, red, blue, and orange for C, O, N, and P atoms), dinucleotide primer (green, red, blue, and orange for C, O, N, and P atoms), RNAP active-center catalytic Mg^{2+}(I) (violet sphere), and RNAP bridge helix (gray ribbon).

B. Contacts of RNAP residues (gray, red, and blue for C, O, and N atoms) with primer and RNAP active-center catalytic Mg^{2+}(I). RNAP residues are numbered both as in *T. thermophilus* RNAP and as in *E. coli* RNAP (in parentheses).

C. Schematic summary of structures. Template-strand DNA (yellow); primer (green); RNAP bridge helix (gray); RNAP active-center catalytic Mg^{2+}(I) (violet). Note that, in contrast to structures with template-strand purine at position TSS-2 (Figure 7), in this structure with template-strand pyrimidine at position TSS-2, no density for CMPcPP is observed.

A



B

replicate	enzymatic treatment	relative efficiency
1	PNK + Rpp	0.91
	Rpp	0.91
	PNK	0.76
	none	1
2	PNK + Rpp	0.90
	Rpp	0.92
	PNK	0.75
	none	1
3	PNK + Rpp	0.92
	Rpp	0.98
	PNK	0.75
	none	1

Figure S8. Primer-dependent initiation *in vivo*: MASTER data analysis

A. 5'-ppp distributions (top panels) and 5'-OH distributions (bottom panels) calculated using uncorrected read counts (left) or using read counts computed using correction factors that account for inefficiencies of enzymatic processing (corrected, right).

B. Relative enzymatic processing efficiencies computed for each replicate.

Table S1. Primer-dependent initiation in natural, chromosomally-encoded *E. coli* promoters

promoter	TSS-5 to TSS+4	% 5'-OH (mean ± SD; N =3)
bhsA	TAGTTaTCGAT	88.4 ± 2.0
btuR	TCTTTaTGGTT	76.6 ± 6.6
yfeC	GTGTTaTAAAA	54.0 ± 9.0
yedY	CTTCTaTGAAT	45.8 ± 2.0
rssB	AAGTTaATTCT	44.8 ± 7.3
tomB	AAGCTaACAAA	43.2 ± 3.0
qmcA	TGTTTaATCAA	38.1 ± 1.4
micL	TCATTaTGGCC	37.8 ± 5.1
ytfK #1	TCTTTaTCAAG	36.9 ± 10.6
ygaV	TACTTaATTTA	36.8 ± 3.0
ybhL	AAGATaTTCCT	36.1 ± 1.8
fadD	ATGTTaACGGC	34.1 ± 2.3
yjjZ	TGGTTaTCATT	30.6 ± 2.1
yejG	GATGTaAACAA	28.3 ± 1.6
rraB	AGAGTaTCGGC	26.7 ± 1.7
hipB	ATAATaTCCCC	26.3 ± 3.6
rlmE #1	CATATaTCACC	25.0 ± 6.5
yibF	GTTATaTCTAA	24.0 ± 12.2
maeB	GTGTTaTAGGA	23.9 ± 4.8
csiD	AGATTaTGGCT	23.1 ± 1.2
purA	TTTTTaAGCAA	22.6 ± 0.1
ftsB	AGATTaTGTTC	20.6 ± 1.8
rodZ	ATGATaACGGT	19.4 ± 7.3
ycbK	TAATTaGCATG	18.7 ± 10.3
ycgN	GTTATaAGGTG	18.7 ± 1.3
uspE	TAATTaCAAAC	18.7 ± 0.4
htpX	TGGGTaTCGCA	18.5 ± 2.8
baeR	GAGTTaCCGCT	17.6 ± 2.4
csrA #2	TGGTTaGCGAG	15.9 ± 1.4
araC	TTGTTaCGCGT	15.8 ± 5.8
melR	CTGATaTTCCA	15.3 ± 4.0
bssS	GGTATaCTGAA	15.2 ± 3.3
ygiB	TGGGTaTTTAC	15.0 ± 0.7
frdA	TACCTaTAAAG	14.0 ± 2.5
ynfM	TGTATaAGCCT	13.9 ± 2.5
ppiB	ACCCTaTATAT	13.5 ± 5.2
lptD	GCATTaTCATA	13.4 ± 2.3
nrdH	TCTGTaTCAAC	12.8 ± 3.2
chaC	CCGTTaGTGGT	12.5 ± 2.1
ycbB	AAGATaAGCCT	12.3 ± 2.3
yabI	TTGGTaACGAA	12.2 ± 1.0
clpP #1	GAAATaTGGTG	12.1 ± 0.7
rpoD #1	CAGATaAGAAT	11.6 ± 0.6
gadY	AACTTaCTGAG	11.5 ± 3.5
rapA	TACGTaTGGAC	11.3 ± 2.2
sufA	CTGTTaTACGC	11.2 ± 0.8
rnlB #2	TTTTTaATGTG	11.1 ± 0.4
ubiC	CCTTTaCGTTA	10.4 ± 1.5
yniB #1	TGTGTaAGGTA	9.5 ± 3.7

gpmA	TGCGTaAGCAT	9.5 ± 2.7
rob	CAATTaCCTGA	9.0 ± 1.0
sseA	AGCCTaACCCC	9.0 ± 3.2
pgk	ACTATaCGGAA	9.0 ± 1.5
bamE	GGCATaTTACA	8.6 ± 1.2
yebW	ACGGTaCGGCA	8.1 ± 2.8
mntS #2	TATGTaATGCA	8.1 ± 0.7
mntH	AACATaGCAAA	8.0 ± 0.7
exuR	AATATaGTCTG	7.1 ± 1.0
yfgG	TAAGTaCCGTC	7.0 ± 0.9
chpS	TGTATaACTTA	6.9 ± 0.8
fumA	TTGTTaCTCGC	6.7 ± 2.1
rnlB #1	AAAGTaAGATG	6.5 ± 0.6
gcl #1	TAGATaAAGAA	6.5 ± 0.2
ydcJ	AGATTaCTGAC	6.2 ± 0.6
moaA	GGCATaACCAC	6.0 ± 1.1
zapB	GAGCTaACTAA	5.3 ± 2.8
fabG	CCGGTaTCACT	5.2 ± 0.8
bsmA	TGGTTaGCAGG	5.2 ± 1.1
sibC	TGAGTaAGGGT	5.2 ± 1.0
zapA	CAAATaCTGAA	5.0 ± 0.8
uspB	TCTATaGAGCC	4.7 ± 0.4
ypfM #1	CCCATaATTAT	4.5 ± 1.3
ytfB	GCAATaTCCCG	4.3 ± 1.4
mtlA	GCAGTaTCTAC	4.1 ± 2.5
mqsR	AAAGTaACAGG	4.0 ± 0.1
yobB	TGAGTaTTTTC	3.4 ± 0.9
ligT	GCGGTaATTCA	3.2 ± 0.9
yobF	AAAATaCGCCA	3.2 ± 1.7
ygaC	AAGATaCGGCG	2.2 ± 3.8
yccX	AACGTaCACTC	1.6 ± 0.7
ihfB	GCACTaAGGGC	1.2 ± 2.0
pgpA	GTCATaCTTCA	1.1 ± 1.2
cspE #1	AAGGTaACGTT	1.0 ± 1.7
yeaY	AGAGTaTGCCC	0.8 ± 1.4
ygdR	TAGGTaGGCCA	0.8 ± 0.5
dcuA	GTTGTaGAACT	0.8 ± 0.7
aldB	GTTATaCCTCA	0.7 ± 0.2
aspA	CGGGTaTTCGG	0.6 ± 0.6
cspD	TTTCTaGAGTT	0.5 ± 0.6
slyD	TGAGTaCACGG	0.4 ± 0.7
ryeA	ACTATaAAGTC	0.2 ± 0.2
chiX	GAGTTaCACCG	0.1 ± 0.1
obgE	TATGTaCAATT	0.0 ± 0.0

Table S2. Crystal structure statistics.

	RPo[$A_{TSS-2}A_{TSS-1}T_{TSS}$] UpA-CMPcPP	**RPo[$A_{TSS-2}C_{TSS-1}C_{TSS}$] GpG-CMPcPP**	**RPo[$T_{TSS-2}C_{TSS-1}C_{TSS}$] GpG**
Data collection			
Space group	C2	C2	P21
Cell dimensions			
a, b, c (Å)	183.8, 103.4, 295.6	184.1, 103.3, 295.6	185.1, 104.3, 297.0
α, β, γ (°)	90.0, 99.0, 90.0	90.0, 98.9, 90.0	90.0, 98.4, 90.0
Resolution (Å)	50.00-2.80 (2.85-2.80)	50.00-2.90 (2.95-2.90)	50.00-3.40 (3.46-3.40)
R_{sym} or R_{merge}	0.113 (1.027)	0.076 (0.429)	0.128 (0.514)
I/σI	12.7 (1.2)	15.0 (1.7)	9.1 (2.1)
Completeness (%)	95.7 (91.7)	96.6 (86.7)	96.4 (85.2)
Redundancy	3.9 (3.3)	3.5 (2.7)	3.9 (3.4)
$CC_{1/2}$ in highest shell	0.536	0.895	0.850
Refinement			
Resolution (Å)	50.00-2.80	50.00-2.90	50.00-3.34
No. reflections	127031	100013	161620
Rwork/ Rfree	0.203/0.243	0.191/0.252	0.201/0.247
No. of atoms	28596	28630	56841
B-factors ($Å^2$)	72.35	48.76	96.54
R.m.s deviations			
Bond lengths (Å)	0.005	0.009	0.003
Bond angles (°)	0.718	1.067	0.563
Ramachandran plot			
Favored (%)	97.33	93.79	97.64
Allowed (%)	2.67	6.21	2.36
Disallowed (%)	0	0	0
PDB code	7EH0	7EH1	7EH2

Numbers in parenthesis are for the highest resolution

Table S3. Oligonucleotides.

sequence (5' to 3')	name	description
ATATAAGCTTTTGCTTTCTATTTAACTGAAATTT	k413	forward primer to clone -100/+15 of the *bhsA* promoter region. Primer contains a 5' HindIII site.
CCTCTCTGCCGGATCC	k711	binds positions +16 to +31 (relative to the TSS) of *bhsA* RNA; used in primer extension assays of Figure 6B
TATATAGGATCCTACTTAACGATCGATA**C**CTAAATGATAG	k1036	reverse primer to clone p*bhsA* G:C_{TSS-2}
TATATAGGATCCTACTTAACGATCGCTAACTAAATGATAG	k1040	reverse primer to clone p*bhsA* G:C_{TSS+2}
GTTCAGAGTTCTACAGTCCGACGATCGCGGCCGCAGGCTTGACACTTTATGCTTCGGCTCGTATAATGTGTTTATGTGTGAGCGGAGGCGGAGGCCGTT	k1134	*lac*CONS nontemplate strand with $T_{TSS-2}T_{TSS-1}A_{TSS}$
GTTCAGAGTTCTACAGTCCGACGATCGCGGCCGCAGGCTTGACACTTTATGCTTCGGCTCGTATAATGTGTGTATGTGTGAGCGGAGGCGGAGGCCGTT	k1135	*lac*CONS nontemplate strand with $G_{TSS-2}T_{TSS-1}A_{TSS}$
GTTCAGAGTTCTACAGTCCGACGATCGCGGCCGCAGGCTTGACACTTTATGCTTCGGCTCGTATAATGTGT/IDSP/TATGTGTGAGCGGAGGCGGAGGCCGTT	k1139	*lac*CONS nontemplate strand with $X_{TSS-2}T_{TSS-1}A_{TSS}$, where X is an abasic site. /IDSP/ indicates an abasic site.
AACGGCCTCCGCCTCCGCTCACACATAAACACATTATACGAGCCGAAGCATAAAGTGTCAAGCCTGCGGCCGCGATCGTCGGACTGTAGAACTCTGAAC	k1136	*lac*CONS template strand with $A_{TSS-2}A_{TSS-1}T_{TSS}$
AACGGCCTCCGCCTCCGCTCACACATACACACATTATACGAGCCGAAGCATAAAGTGTCAAGCCTGCGGCCGCGATCGTCGGACTGTAGAACTCTGAAC	k1137	*lac*CONS template strand with $C_{TSS-2}A_{TSS-1}T_{TSS}$
AACGGCCTCCGCCTCCGCTCACACATA/IDSP/ACACATTATACGAGCCGAAGCATAAAGTGTCAAGCCTGCGGCCGCGATCGTCGGACTGTAGAACTCTGAAC	k1138	*lac*CONS template strand with $X_{TSS-2}A_{TSS-1}T_{TSS}$, where X is an abasic site. /IDSP/ indicates an abasic site.
GTTCAGAGTTCTACAGTCCGACGATCGCGGCCGCAGGCTTGACACTTTATGCTTCGGCTCGTATAATGTGTTCATGTGTGAGCGGAGGCGGAGGCCGTT	k1568	*lac*CONS nontemplate strand with $T_{TSS-2}C_{TSS-1}A_{TSS}$
AACGGCCTCCGCCTCCGCTCACACATGAACACATTATACGAGCCGAAGCATAAAGTGTCAAGCCTGCGGCCGCGATCGTCGGACTGTAGAACTCTGAAC	k1570	*lac*CONS template strand with $A_{TSS-2}G_{TSS-1}T_{TSS}$
GTTCAGAGTTCTACAGTCCGACGATCGCGGCCGCAGGCTTGACACTTTATGCTTCGGCTCGTATAATGTGTTTGTGTGTGAGCGGAGGCGGAGGCCGTT	k1569	*lac*CONS nontemplate strand with $T_{TSS-2}T_{TSS-1}G_{TSS}$
AACGGCCTCCGCCTCCGCTCACACACAAACACATTATACGAGCCGAAGCATAAAGTGTCAAGCCTGCGGCCGCGATCGTCGGACTGTAGAACTCTGAAC	k1571	*lac*CONS template strand with $A_{TSS-2}A_{TSS-1}C_{TSS}$

oligo pool for analysis of 93 chromosomally-encoded promoters that use UpA as primer

Sequence (Illumina RT primer sequence is underlined)	name	target promoter
GCCTTGGCACCCGAGAATTCCAAGTGGCCTTATGCAGATGAATGACG	k888	*bhsA*
GCCTTGGCACCCGAGAATTCCATCCGGCATCGAGGAAGTGGTGCAGG	k889	*micL*
GCCTTGGCACCCGAGAATTCCAACCTTTCGGGATGGAAAAAACTTAC	k890	*tomB*
GCCTTGGCACCCGAGAATTCCATGAAAGCTCATCATGTCATACGTCC	k891	*hipB*
GCCTTGGCACCCGAGAATTCCAGGCCTGTAATTGCGCGAGTTCAGTC	k892	*ygaV*
GCCTTGGCACCCGAGAATTCCAATCAGAACGTGGGAATCTGTCCATG	k893	*ybhL*

GCCTTGGCACCCGAGAATTCCATCGATTGGTAACTCGGTCATACTTC	k894	*baeR*
GCCTTGGCACCCGAGAATTCCAAGACAACGTTATTCGAGGTTCAATG	k895	*bssS*
GCCTTGGCACCCGAGAATTCCATTAAAAGCATTCTGTAACAAAGCGG	k897	*araC*
GCCTTGGCACCCGAGAATTCCAGGAAAAACCTCCTGTTGTACCGTCC	k898	*qmcA*
GCCTTGGCACCCGAGAATTCCATAATGAAGTCACTTATTTTCCCCGG	k900	*yejG*
GCCTTGGCACCCGAGAATTCCAATTCTGCAAACCCTATGCTACTCCG	k901	*yabI*
GCCTTGGCACCCGAGAATTCCAACTGCCCAGCGTACGTTGCAACATG	k902	*yjjZ*
GCCTTGGCACCCGAGAATTCCATGCCCAGTACGACGACGTTGTTACC	k905	*purA*
GCCTTGGCACCCGAGAATTCCAAGAAAGGTACATCGCTCATCAGGTG	k907	*ycgN*
GCCTTGGCACCCGAGAATTCCATAGTACGGCTCACTTGAAATCCTTG	k908	*ynfM*
GCCTTGGCACCCGAGAATTCCAGGATAACTCATCGTAACCAATTGCG	k909	*rlmE #1*
GCCTTGGCACCCGAGAATTCCACCAGTTGTTCCGGGTTTGCCATGGC	k910	*rraB*
GCCTTGGCACCCGAGAATTCCAGGCGTGGTTCAGTGATTTCCATATG	k911	*exuR*
GCCTTGGCACCCGAGAATTCCACGCATCATAATTTTCTTTTTACCTC	k912	*htpX*
GCCTTGGCACCCGAGAATTCCAGTGGCGTATGGATTTTGTCCGTTTC	k913	*ygiB*
GCCTTGGCACCCGAGAATTCCACTGGCTTTACTCAATAGTGGCATGC	k914	*rssB*
GCCTTGGCACCCGAGAATTCCAATCAAACATATGGTTGAGTGAGTCG	k916	*rpoD #1*
GCCTTGGCACCCGAGAATTCCATGCATCCTGTTCCGTTTGATTGGTG	k917	*ppiB*
GCCTTGGCACCCGAGAATTCCAACTCAACCCCAGCAGTGACGGGGGC	k918	*ycbK*
GCCTTGGCACCCGAGAATTCCATAAATCGTCAACCATGGTACGCAAC	k919	*csrA #2*
GCCTTGGCACCCGAGAATTCCATCTTTAAAGGCGATATGATAGGCGC	k920	*yniB #1*
GCCTTGGCACCCGAGAATTCCAAAGGTTCCTGAATGCATGTCCATCG	k921	*sufA*
GCCTTGGCACCCGAGAATTCCAAAAGCACCGTATCAGTTGACCCAGG	k922	*moaA*
GCCTTGGCACCCGAGAATTCCACCTGATCCATAAAATATCCTCATCC	k923	*rob*
GCCTTGGCACCCGAGAATTCCAACGCCAATACTCACGAAAGTATGAC	k924	*ycbB*
GCCTTGGCACCCGAGAATTCCACCTCGTCCCGATTACCGGTGACGCC	k925	*zapB*
GCCTTGGCACCCGAGAATTCCATAAAATGCGTGTTCGTCGTCATCGC	k926	*fadD*
GCCTTGGCACCCGAGAATTCCAAAGACCTCCTGACTTGCTAATCCCG	k927	*mntS #2*
GCCTTGGCACCCGAGAATTCCACCAGTGGTGAACGTTGGTAGTCCAG	k928	*rapA*
GCCTTGGCACCCGAGAATTCCAAGTATCTCTGTATCAACAGAGAGAC	k929	*ygaC*
GCCTTGGCACCCGAGAATTCCATATGGACGGATTATCAGCTAGTCCC	k933	*clpP #1*
GCCTTGGCACCCGAGAATTCCAACAATGACGCCCATTTCTTTGCATC	k1201	*yccX*
GCCTTGGCACCCGAGAATTCCAATGTACCACATTTGTCCATTGTTAC	k1202	*ftsB*
GCCTTGGCACCCGAGAATTCCAAACATAACAATTCCCTGAATGTATG	k1203	*yfeC*
GCCTTGGCACCCGAGAATTCCAGCTGCACATAAACGTATCTGTATTC	k1204	*melR*
GCCTTGGCACCCGAGAATTCCAGTGGTGGCTGGTAATTAAGCTGATG	k1205	*btuR*
GCCTTGGCACCCGAGAATTCCAAGCGAGTGCCGCCGATTGGCATTAC	k1207	*ytfK #1*
GCCTTGGCACCCGAGAATTCCAAAACCGTTATAACACTCCCTGTTGG	k1208	*gadY*
GCCTTGGCACCCGAGAATTCCAAAAGAACAGACGTTGCGGTTCAGAC	k1209	*ligT*
GCCTTGGCACCCGAGAATTCCATCCGAATAGCGTCAGTGGTGTTAGG	k1210	*lptD*

GCCTTGGCACCCGAGAATTCCAAAGAATATTTGACTCTCTAATATTG	k1211	*yibF*
GCCTTGGCACCCGAGAATTCCATCAGGCTACCAGCATTAGCGGCGGG	k1212	*yedY*
GCCTTGGCACCCGAGAATTCCACCGACTGCAGAAATTATTACTGCCC	k1213	*ygdR*
GCCTTGGCACCCGAGAATTCCAGAGCATGGGAAGAATAGTGGCGCAG	k1214	*slyD*
GCCTTGGCACCCGAGAATTCCAGGTCTCTTTTTATCTGTTAAAAGCC	k1215	*ryeA*
GCCTTGGCACCCGAGAATTCCAGCAAACAAAAAGTACCAGCGCGAAC	k1216	*yebW*
GCCTTGGCACCCGAGAATTCCAGCTAAGCTGCGCTATCACTCCAGGC	k1217	*obgE*
GCCTTGGCACCCGAGAATTCCAAAGTACCTGAAAGTTACGGTCTGCG	k1218	*frdA*
GCCTTGGCACCCGAGAATTCCAGTTCCACTCTTCATCTTTCATTCGC	k1219	*fabG*
GCCTTGGCACCCGAGAATTCCACGAGGTTTTTTGTTAATGTGGCGGC	k1220	*mtlA*
GCCTTGGCACCCGAGAATTCCAGCGCTTATTAACAGTCAGTCTCAGG	k1221	*sibC*
GCCTTGGCACCCGAGAATTCCATGTGGTTTCGAATGTTTTTCGACCG	k1222	*aspA*
GCCTTGGCACCCGAGAATTCCAGTAAGTGACTGGGGTGAACGAATGC	k1223	*pgpA*
GCCTTGGCACCCGAGAATTCCAAACGCTATGATGTCCGTGGTAAACC	k1224	*ytfB*
GCCTTGGCACCCGAGAATTCCATGCCTGACTCACAAAAGGTTCCTTG	k1225	*yfgG*
GCCTTGGCACCCGAGAATTCCAAAGTTCCAAAACCTGCCGGCTGCTG	k1226	*rnlB #1*
GCCTTGGCACCCGAGAATTCCAATCCTCTTATGAGATGTAGGGTGAC	k1227	*csiD*
GCCTTGGCACCCGAGAATTCCACTTCCGGAGTAATGAAACCGAATCC	k1228	*cspE #1*
GCCTTGGCACCCGAGAATTCCATGATAATAATTCTCATTATATTGCC	k1229	*gpmA*
GCCTTGGCACCCGAGAATTCCATGCTGAAGGGGGATTATTGGTCATG	k1230	*aldB*
GCCTTGGCACCCGAGAATTCCAATGACTTCCTTTTCAAAATGACTGC	k1231	*yobB*
GCCTTGGCACCCGAGAATTCCACTTTGAGCAAGTCCAAACTCTCACC	k1232	*maeB*
GCCTTGGCACCCGAGAATTCCAGGTTTGGTTATGCTCTGGGCGGGTG	k1233	*fumA*
GCCTTGGCACCCGAGAATTCCATAACTCATAGCTGATATTCAATGCG	k1234	*zapA*
GCCTTGGCACCCGAGAATTCCAATAACATTCTTTTGAACCGCCATGG	k1236	*yeaY*
GCCTTGGCACCCGAGAATTCCACACTGTGCGGATCGTGGTTAAAATC	k1237	*pgk*
GCCTTGGCACCCGAGAATTCCACTCATCCGCCGTGATGCTGTTCGCC	k1238	*ydcJ*
GCCTTGGCACCCGAGAATTCCAAAATGAAGATACGGCGCGATGATAC	k1239	*nrdH*
GCCTTGGCACCCGAGAATTCCAGCTGGCATCGACCACAGTTTCCATG	k1241	*rnlB #2*
GCCTTGGCACCCGAGAATTCCACACCAATAATAAACTGGCAAACCGG	k1242	*bsmA*
GCCTTGGCACCCGAGAATTCCAGCACTGTTCCCCCATCTTTTTATGG	k1243	*chpS*
GCCTTGGCACCCGAGAATTCCAATATCATTGTGCTTAACCTTGCCAG	k1244	*yobF*
GCCTTGGCACCCGAGAATTCCATGAACCATCATGCTCACCCACACCG	k1245	*rodZ*
GCCTTGGCACCCGAGAATTCCAGACTCAAACGTGTATGTGGTGTGCG	k1246	*mqsR*
GCCTTGGCACCCGAGAATTCCACTTTCCAGTTCCCCAGTTCACGTTC	k1247	*ypfM #1*
GCCTTGGCACCCGAGAATTCCAGACAACTTTTTGGTGTCTTCCGGAC	k1248	*chaC*
GCCTTGGCACCCGAGAATTCCATGGAATAAACCGGGGTCAGAGAGGG	k1249	*cspD*
GCCTTGGCACCCGAGAATTCCACCAGGTCCTGGAGAAACCGCTTTTG	k1251	*uspB*
GCCTTGGCACCCGAGAATTCCACAGTTGCGTTAACGCGGGGTGTGAC	k1252	*ubiC*
GCCTTGGCACCCGAGAATTCCATCATCGTAAATTGCTGAAGCTATGC	k1253	*gcl #1*

GCCTTGGCACCCGAGAATTCCATGATACATAGCCATACAGGGTCTCC	k1254	*uspE*
GCCTTGGCACCCGAGAATTCCAATAGTGATTTGATTCCTTTTCGGGC	k1258	*bamE*
GCCTTGGCACCCGAGAATTCCAACTCTCAACGCGATAGTTCGTCATC	k1259	*mntH*
GCCTTGGCACCCGAGAATTCCACAAATTAAGCGGCTGCTGTTGCTGC	k1260	*ihfB*
GCCTTGGCACCCGAGAATTCCAGGAATTTCATTTTTTTATTATTATG	k1262	*chiX*
GCCTTGGCACCCGAGAATTCCAGGCTCCTACAAACCATGTCGTGGAC	k1264	*sseA*
GCCTTGGCACCCGAGAATTCCATCCCCCCAATCTGGCGCCCAAGAAG	k1265	*dcuA*