Article

Control of transcription initiation by biased thermal fluctuations on repetitive genomic sequences

Masahiko Imashimizu[1*], Yuji Tokunaga[1], Ariel Afek[2], Hiroki Takahashi[3,4,5], Nobuo Shimamoto[6] and David B. Lukatsky[7*]

[1]Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology, Tokyo, 135-0064, Japan
[2]Center for Genomic and Computational Biology, Department of Biostatistics and Bioinformatics, Duke University.
[3]Medical Mycology Research Center, Chiba University, Chiba 260-8673, Japan
[4]Molecular Chirality Research Center, Chiba University, Chiba 263-8522, Japan.
[5]Plant Molecular Science Center, Chiba University, Chiba 260-8675, Japan
[6]National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan.
[7]Department of Chemistry, Ben-Gurion University of the Negev, Beer-Sheva, Israel; Ben Gurion Blvd 1, Beer-Sheva 8410501
* Correspondence: D.B.L.: Tel: +972-8642-8370; E-mail: lukatsky@bgu.ac.il; M.I.: Tel: +81-3-3599-8232; E-mail: m.imashimizu@aist.go.jp

Abstract: In the process of transcription initiation by RNA polymerase, promoter DNA sequences affect multiple reaction pathways determining the productivity of transcription. However, the question of how the molecular mechanism of transcription initiation depends on sequence properties of promoter DNA remains poorly understood. Here, combining the statistical mechanical approach with high-throughput sequencing results, we characterize abortive transcription and pausing during transcription initiation by *Escherichia coli* RNA polymerase at a genome-wide level. Our results suggest that initially transcribed sequences enriched with thymine bases represent the signal inducing abortive transcription. On the other hand, certain repetitive sequence elements broadly embedded in promoter regions constitute the signal inducing pausing. Both signals decrease the productivity of transcription initiation. Based on solution NMR and in vitro transcription measurements, we also suggest that repetitive sequence elements of promoter DNA modulate the rigidity of its double-stranded form, which profoundly influences the reaction coordinates of the productive initiation via pausing.

Keywords: promoter sequences; repetitive sequences; pausing; abortive initiation; RNA polymerase; dsDNA rigidity

1. Introduction

In bacteria, transcription at a promoter is initiated by σ factor that forms a holoenzyme by binding to RNA polymerase (RNAP) core enzyme. The principal σ factor in *Escherichia coli* is termed σ^{70}. *E. coli* promoters targeted for transcription initiation by σ^{70} holoenzyme have been characterized by having two consensus motifs approximately 10 and 35 bases upstream of the transcription start site (TSS). These motifs consist of a TATAAT (-10 box) and a TTGACA (-35 box), conserved in the promoters with high binding affinity to σ^{70} holoenzyme [1,2]. However, biologically functional promoters with high transcriptional activities usually do not have the full consensus motifs but rather have non-local sequence signatures across the overall promoter region [3]. The reason for this has been a long-standing puzzle with respect to the regulatory mechanism of transcription initiation.

In transcription initiation, σ^{70} holoenzyme synthesizes non-productive (abortive) short RNA [4,5]. When it occurs in a promoter, this process is much slower than the productive initiation and

the following elongation processes [6,7]. During abortive synthesis, a ternary initiation complex of the σ^{70} holoenzyme starts transcription and then backtracks to shorten the RNA-DNA hybrid, thereby releasing short RNAs [8-11]. Such complex is termed moribund complex (see review [12]). Initiation pathway leading to the abortive synthesis by the moribund complex is branched from the pathway leading to full-length RNA synthesis by the productive complex [6], which can be a mechanism controlling transcription initiation by RNAPs of *E. coli* and other bacteria [13-15] (Figure 1). The binary moribund complex can be converted into the productive complex by binding of allosteric effectors like Gre proteins to the complex [16]. This function of Gre proteins is different from their well-known function, i.e. cleavage of the 3′ RNA that is extruded from the active center of the backtracked polymerase [17]. The level of abortive initiation depends on sequences both the upstream and the downstream of TSS, and the first ~20 bp of the downstream sequence is termed initially transcribed sequence (ITS) [18,19].

On the other hand, pausing that occurs on a pathway of productive initiation can delay transcription kinetics on a physiological timescale by affecting promoter escape [8-10,20]. It has been reported that specific sequences in ITS can induce pausing during initiation [19]. Therefore, not only (i) the fraction of the moribund complex that is generated and branched from the entire binary complex fraction but also (ii) the lifetime (and/or frequency) of pausing in the productive complex may constitute the sequence-specific mechanisms controlling transcription during initiation. To date, individual sequence signals that are responsible for abortive synthesis and pausing have not been separately identified.

Previously, we identified a highly conserved sequence motif that induces elongation pausing in *E. coli* [21]. This motif impedes forward translocation of RNAP, as well as the following NTP addition [21]. However, our later analysis revealed that the presence of the conserved sequence motif alone is not solely responsible for RNAP pausing [22]. In particular, we demonstrated that, during elongation pausing, repetitive sequence elements can increase the magnitude of diffusive backtracking of RNAP on the DNA upstream of the pausing site, generating a large variation in the lifetimes of RNAP pausing under the catalytic control by the conserved sequence motif [22]. Therefore, our approach allowed global prediction of elongation pausing in *E. coli*.

In this study, using a similar approach, we characterized abortive transcription and pausing during initiation in *E. coli* at a genome-wide level. Our results suggest that T-rich signal located in ITS can be the signal inducing abortive synthesis, while repetitive sequences of any base types widely distributed in promoter regions can be the signal inducing pausing. We also identify the rigidity of double-stranded DNA (dsDNA) as a possible physicochemical origin affecting reaction coordinates of the productive initiation via pausing.

Peer-reviewed version available at Biomolecules 2020, 10, 1299; doi:10.3390/biom10091299

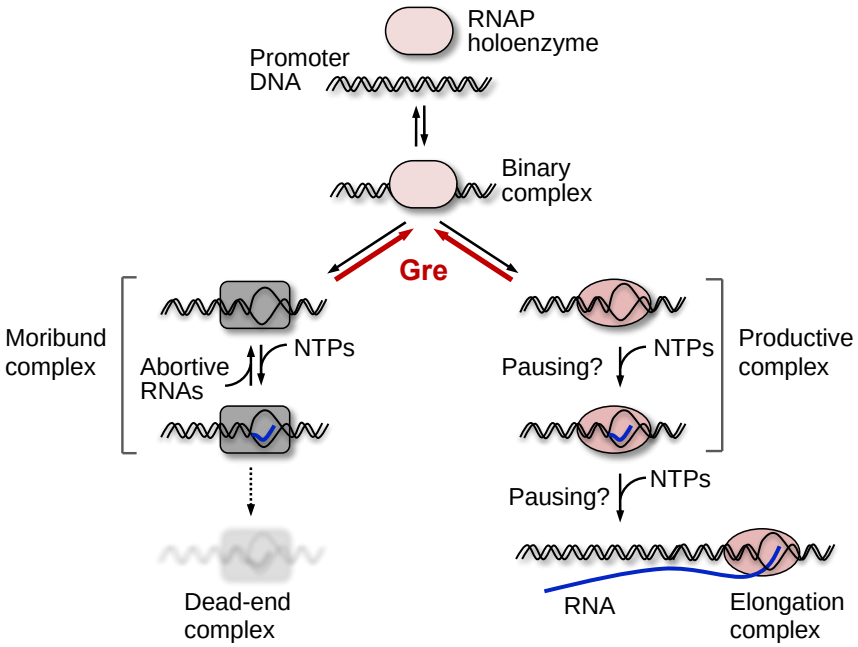


Figure 1. Branched initiation pathway. The action of Gre proteins at branches is shown by red color. In the presence of Gre proteins, the branching becomes reversible (thick red arrows) so that the moribund and the productive complexes can be exchanged each other [13]. Abortive RNA synthesis by the moribund complex is a slow process compared to full-length RNA synthesis (usually, up to 20 min [7,20]), which is reduced by Gre proteins, and thus is genome-widely detectable by RNET-seq with Gre-dependency of the data [21]. At several promoters, the moribund complex is further converted into a dead-end complex that still retains abortive RNA but has no elongation activity [6]. RNAP also pauses during productive initiation [8-10,20], which often involves backtracking of RNAP by one bp and thus can be reduced by Gre. Long-lifetime pausing by the moribund complex is incorporated in the processes of abortive transcription.

2. Results

We have previously developed RNase-footprinting followed by NET-seq (RNET-seq) method to identify the complexes that were paused during transcription elongation in *E. coli* wild type (WT) and in an isogenic strain deficient in genes for GreA and GreB (*ΔgreAB*) [21]. Briefly, *E. coli* cells were rapidly lysed and any transcribing RNAPs were released from the genomic DNA and co-transcriptional translation by digestion with DNase I and RNase A, respectively. All RNAPs including those associated with the fragmented dsDNAs and their 5′-truncated nascent RNAs were immobilized on Ni^{2+}-NTA beads via the histidine-tagged β′ subunit and then washed. The 5′ ends of the transcripts in the ternary complexes were trimmed with RNase T1/V1 to leave a minimal length of RNA protected by RNAP. The RNases were removed by further washing. Elution with imidazole generated ternary complexes carrying ~6-30 nt long transcripts.

In RNET-seq method, the longer the time that RNAP occupies a particular DNA site during elongation, the stronger pausing at the DNA site are detected. Here we noticed that this method can also collect 6~13 nt long abortive RNA transcripts retained (prior to their release) in the moribund complex, in addition to RNA transcripts retained in the paused productive complex. The 5′ end of the unreleased abortive transcripts should be mapped at TSS and increased in the *ΔgreAB* cells as compared with WT cells (Figure 1). To confirm this possibility, we mapped and aligned short RNET-seq reads to TSS and the close vicinity in 775 σ^{70} promoter sequences that are experimentally identified and are available from RegulonDB [23] (Figure 2A). Figure 2B shows the results for 7~10 nt transcripts for the 775 promoters.

We then classified those σ^{70} promoters into the following three groups according to the magnitude of the ratio, *X*, of the amounts of nascent RNA transcripts (nrt) in *ΔgreAB* cells,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 August 2020 doi:10.20944/preprints202008.0548.v1

Peer-reviewed version available at Biomolecules 2020, 10, 1299; doi:10.3390/biom10091299

nrt(*ΔgreAB*), to that in WT cells, *nrt*(WT), respectively, X=*nrt*(*ΔgreAB*) /*nrt*(WT). We always define this ratio, X, separately for each transcript length, 7nt, 8nt, 9nt, and 10nt, respectively. We term the three groups as increased ratio ($X \geq 2$), similar ratio ($0.5 < X < 2$), and reduced ratio ($X \leq 0.5$), respectively. Since Gre factors unlikely affect binding of RNAP holoenzyme to specific promoter sequences [17,24], we assume that only the first promoter group ($X \geq 2$) possesses much abortive synthesis, while the second ($0.5 < X < 2$) and third ($X \leq 0.5$) groups possess little abortive synthesis. Transcripts belonging to the third group may originate due to indirect influence or unknown functions of Gre factors. In other words, we suggest that abortive synthesis is predominately represented by the first group, but we do not suggest that all the transcripts in this group are abortive transcripts. As we mentioned above, pausing during productive initiation is also classified in the first group when the pausing involves backtracking. Thus, we hereafter term the first group *abortive/pausing-enriched* group.

Next, we investigated the group-specific sequence properties in terms of the following two different binding modes: (i) specific RNAP-DNA binding on consensus DNA motifs and (ii) nonspecific RNAP-DNA binding on repetitive DNA sequence elements. We assume here that these two types of binding mechanisms are entirely decoupled, i.e. the specific binding mechanism (i) does not affect the nonspecific binding mechanism (ii), and vise versa. Hereafter, we term the former mechanism as *consensus mode* of RNAP-DNA binding, and the latter mechanism as *nonconsensus mode* of RNAP-DNA binding, respectively. The consensus mode conventionally assumes a single (or a few) dominant conformation(s) in the complex. This effect is often represented by information content, the level of sequence conservation within the motif defined [25]. The nonconsensus mode assumes many conformations of the complex that are exchanged as a result of thermal fluctuations. This effect can be modeled as one-dimensional diffusion of RNAP on DNA induced and biased by repetitive DNA sequence elements [26]. In particular, in our recent works we have developed statistical mechanical modeling approach taking into account the effect of certain repetitive DNA sequence elements on protein-DNA binding free energy [27,28]. We have shown in these works that certain repetitive nonconsensus genomic background sequences surrounding a consensus motif can significantly modulate binding of the target protein to DNA via the entropy dominated mechanism [27,28]. We have quantitatively characterized this mechanism using an equilibrium statistical mechanics model without fitting parameters, where actual genomic DNA sequences constitute the only input parameter [27,28]. This statistical concept has allowed us to quantitatively characterize microscopic heterogeneity of protein-DNA complexes stemming from thermal fluctuations as entropy-dominated free energy, which strongly depends on certain repetitive DNA sequence elements recognized by a protein [27-29]. We term this *free energy index for the nonconsensus mode of protein-DNA binding* (FEINC). Using this approach, we have previously predicted that repetitive genomic sequences significantly enhance RNAP pausing during elongation by increasing the number of the paused complex conformations induced by thermal fluctuations [22]. Such a prediction was experimentally verified as the observation of enhanced diffusive backtracking of *E. coli* RNAP in genomic pause sites that are enriched with repetitive sequence elements [22].

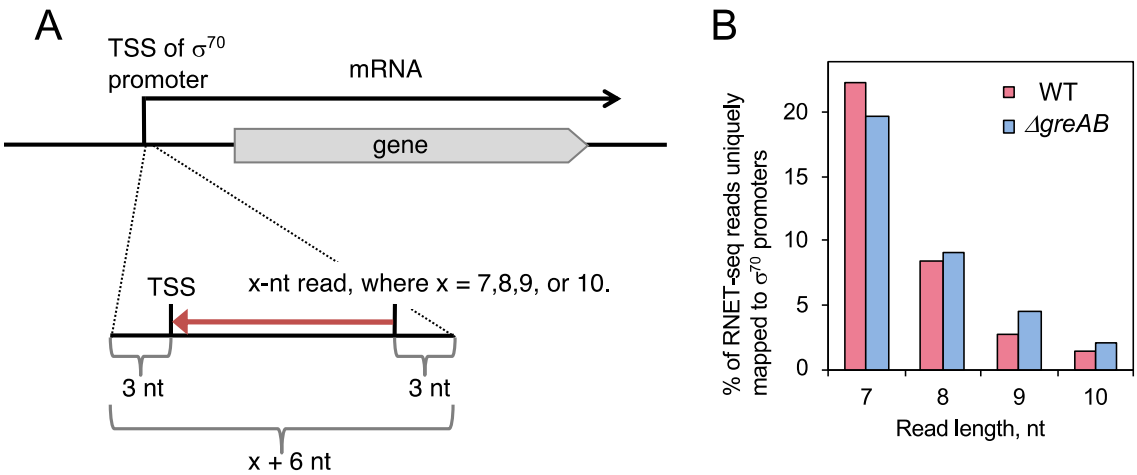


Figure 2. RNET-seq analysis for abortive transcription and pausing during initiation. (A) The short RNET-seq reads of a fixed length (7, 8, 9 or 10 nt long) were mapped to the TSS downstream of 775 σ^{70} promoter regions. We allowed $\pm$3 nt positional fluctuations of TSS. Since the 7-10 nt reads were too short to be uniquely and precisely mapped to the entire *E. coli* genome, we extracted the TSS downstream sequences ($\pm$3 nt) from the genome as reference sequences enabling us to uniquely map these fixed-length reads to the reference. Using this procedure, the 775 experimentally identified σ^{70} promoters were selected from the total of 1873 candidates provided by RegulonDB [23]. (B) The short reads of each length, with sense orientation to mRNA genes, were mapped to the special references by Blat program [30]. The uniquely mapped reads with perfect matches were selected for the analysis performed in the present study. The reads were obtained by RNET-seq of the nascent RNAs of *E. coli* WT and *ΔgreAB* cells [21].

2.1. The significance of consensus mode of RNAP-DNA binding

We found that only one of the 775 promoters (*metY* gene) has full consensus motifs -10 (TATAAT)/-35 (TTGACA). In any groups, the -10 and -35 motifs are not well conserved (information contents <0.5 bits) through the complexes carrying the four different lengths of RNA (Figure 3). This indicates that -10/-35 motifs alone are insufficient to describe the mechanism to initiate transcription at σ^{70} promoters in vivo. In fact, we detected no major differences in -10 or -35 motifs among the three groups of *nrt*(*ΔgreAB*) /*nrt*(WT) ratios (Figure 3). Only in the relatively well-conserved -10 motifs, we observed a minor difference at -7 positions among those three groups: T base at the position -7 tends to be avoided in the abortive/pausing-enriched group (Figure 3).

On the other hand, we found that approximately 6 consecutive T bases are slightly conserved in ITS in the abortive-pausing-enriched group (Figure 3). It has been reported that T bases in ITS stimulates abortive synthesis or pausing during initiation via biasing translocation equilibrium of RNAP toward the pre-translocated state [18,19]. The bias to the pre-translocated state increases probability of backtracking of the RNAP relative to RNA-DNA hybrid, thereby being able to induce abortive transcription. Backtracking can be also induced by the unstable U-dA base pairs within the RNA-DNA hybrid that is encoded by the consecutive T bases [31-33]. Therefore, T-rich ITS signal could be a candidate of a sequence signal to induce a process of abortive transcription in vivo, although it is not clear whether this signal is included in the consensus RNAP-DNA binding effect. More details about the relation between T-rich ITS and RNAP backtracking are discussed below.

Note that the consensus -10/-35 motifs are much better conserved when those promoter DNAs are isolated solely by the binding strength specific to the holoenzyme [34]. Taken together, the -10/-35 motifs undoubtedly determine the strength of the holoenzyme-promoter-DNA binding but are not enough to determine the fate of transcription initiation including abortive synthesis and pausing in vivo. The consecutive T repeats within ITS may participate in inducing abortive synthesis or pausing.

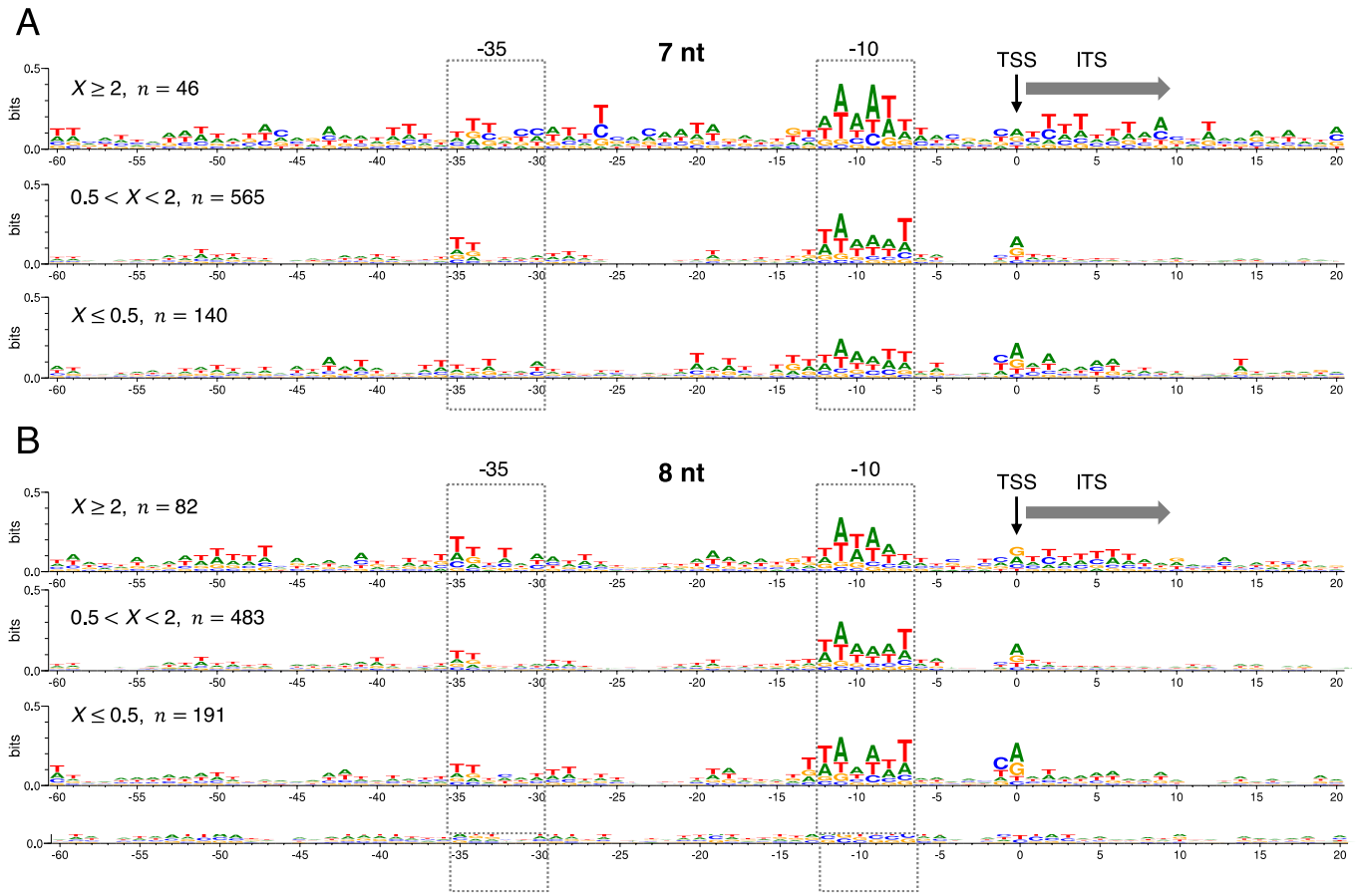


Figure 3. The effect of consensus mode of holoenzyme-DNA binding on initiation complexes having nascent RNAs of 7 nt (A), 8 nt (B), 9 nt (C), and 10 nt (D), respectively. The consensus effect is classified into the three promoter groups where abortive synthesis or pausing was decreased ($X \geq 2$) (top), unaffected ($0.5 < X < 2$) (middle) and increased ($X \leq 0.5$) (bottom) by Gre proteins. Here X represents *nrt*(*ΔgreAB*)/*nrt*(WT) ratio for the RNET-seq reads of each length that is mapped to the close vicinity of TSS of σ^{70} promoters; n represents the number of promoters composing the group. We have excluded rRNA promoters and promoters for unexpressed genes from these 775 σ^{70} promoters. The sequence conservation (information content, bits) in the promoter DNA region (from -60 to +20, where TSS is +1) is represented by Sequence Logo [25]. -10 and -35 motifs are shown by boxes. TSS and ITS are shown by a vertical arrow and by a lateral arrow, respectively.

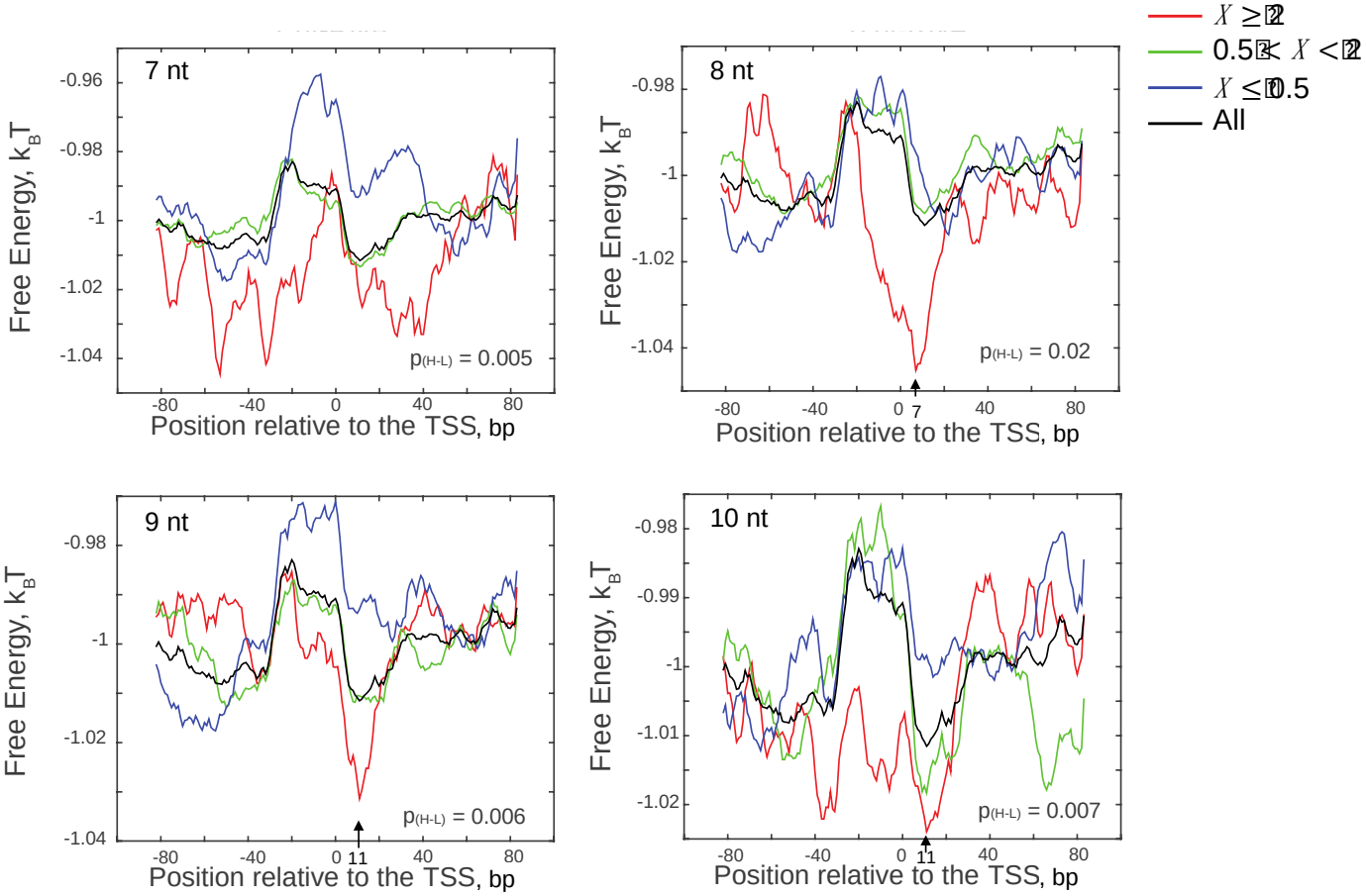


Figure 4. The free energy index for the nonconsensus mode of RNAP-DNA binding (FEINC) was reduced for the initiation complex in the vicinity of TSS. Here we grouped the sequences into three groups according to the ratio, *X*=*nrt*(*ΔgreAB*)/*nrt*(WT), for the RNET-seq reads with the length of 7 nt, 8 nt, 9 nt, and 10 nt, respectively, mapped to TSS of each σ^{70} promoter. Free energy for the nonconsensus binding was calculated as described in [22]. In the calculation of the nonconsensus RNAP-DNA binding free energy, we used the sliding window width, *L* = 30 bp, and we assumed that RNAP-DNA contact window length, M = 8 bp. In each plot for 8 nt, 9 nt, and 10 nt, a base position that has the lowest free energy is shown with an arrow.

2.2. The significance of nonconsensus mode of RNAP-DNA (RNA-DNA hybrid) binding

The sequence effect represented by the average FEINC was significantly different between the abortive/pausing-enriched group (*X*≥2) and the abortive/pausing-depleted group (*X*≤0.5) (Figure 4). The former group (*X*≥2) is characterized by the low FEINC for 7~10-nt transcripts. Interestingly, in the complexes retaining 8-nt-to-10-nt nascent RNAs, the FEINC of the former group (*X*≥2) is the lowest at the site with the position shifted towards the 3' end of the nascent RNAs. This FEINC landscape appears to indicate an enhanced sliding of RNAP along DNA (i.e. backtracking relative to the RNA-DNA hybrid in the case of the ternary complex), according to our previous characterization of the elongation pausing [22]. Likewise, prevention of backtracking is predicted from the high average FEINC (i.e. high energy barrier to backtracking) upstream of TSS in the opposite group (X≤0.5) for the complexes retaining 7-nt or 9-nt nascent RNA (Figure 4) [22].

The RNAP backtracking has been proposed for abortive RNA release form the secondary channel (12). The model is so far consistent with the biochemical/biophysical results obtained by various research groups [8-10,12]. As mentioned above, backtracking is also predicted by the T-rich signal in ITS [18,19]. In fact, slight enrichment of the homopolymeric T in ITS is observed only in the abortive/pausing-enriched group (Figure 3, *X*≥2). Examples of the FEINC for 9 individual promoters are shown in Figure 5. We selected those examples from the three representative groups of DNA

sequences. The first and second group corresponds to $X \geq 2$, with and without the T-rich signal in ITS, respectively, and the third group corresponds to $X \leq 0.5$ (Figure 5). The first and second group exhibit a qualitatively similar FEINC landscape within -40 bp to +40 bp (a valley in the region (0, +20) bp around TSS is observed). Such low FEINC stems from the presence of repetitive sequence elements in the $X \geq 2$ group. Thus, repetitive sequences can accelerate reaction pathways via backtracking of RNAP on DNA, including abortive synthesis and pausing. Higher FEINC within the region -40 bp to +40 bp in the $X \leq 0.5$ group are obtained by less repetitive sequence elements.

We then question (i) whether both the enrichments of >40 bp repetitive sequences and ~6-bp T-rich ITS encode the same signal allowing backtracking, and (ii) whether such respective sequence signals affect differently abortive synthesis and pausing. In order to address these questions, we performed single-round transcription assay using purified *E. coli* RNAP, GreAB proteins and linear 140-bp DNA templates for 9 promoters. The results were analyzed by high-throughput sequencing of total RNA. The sequences of the 9 promoters are shown in Figure 5B: 3 of 6 promoters of the abortive/pausing-enriched group ($X \geq 2$) possess the T-rich ITS, while the other 3 do not possess the T-rich ITS. The remaining 3 promoters belong to the opposite group ($X \leq 0.5$). By 20 min incubation with 100 μM NTPs, more 6-13-nt transcripts were produced by the presence of T-rich ITS in the former group, and the number of those short transcripts were reduced by GreAB, as expected for abortive initiation (Figure 6). Furthermore, the fraction of 6-13-nt transcripts increased time-dependently as predicted from the branched initiation pathway where slow abortive synthesis continues to occur long after completion of full-length RNA synthesis (Figure 1 and Supplementary Figure S1). Thus, our in vitro analysis showed that T-rich ITS induces abortive transcription. We discuss more details about these in vitro results in Supplemental Information.

We further explored a relation between FEINC and transcription initiation from the 9 different promoters in vitro. We found that the FEINC of these promoters tends to positively correlate with the number (i.e. the read count) of long >14-nt transcripts, when pqiA promoter is exclude as an outlier (Figure 7A). The opposite (i.e. negative correlation) trend was observed in the relative fraction of short, 6-13-nt transcripts (Figure 7B). In order to obtain such correlations, we added Gre proteins and shortened the reaction time (1.5 min) (Figure 7B). Under these conditions, the abortive synthesis by moribund complex (typically >1.5 min) should be decreased or negligible, since Gre proteins enable switching of the complex into the productive complex by lowering the high activation energy to the pathway branched in these two complexes (Figure 1 and Supplementary Figure S1) [12,13]. Thus, we interpreted the low FEINC as indicating pausing of the productive complex rather than abortive synthesis of the moribund complex. The exceptional behavior of pqiA promoter appears to be consistent with its belonging to $X \leq 0.5$ group, especially for the high FEINC over -40 bp to +40 bp. It is likely that the high FEINC may contribute to generate the exceptionally high activation energy to the branched pathway, making it less dependent on Gre proteins.

The pausing of the productive initiation complex likely stems from diffusive backtracking of RNAP as shown in elongation [22], although its detection in the presence of Gre proteins appears to be inconsistent with backtracking. Compared to the elongation complex, the reduction of pausing lifetime by the Gre-dependent 3′ RNA cleavage in the initiation complex may be limited because the time that the 3′ RNA end dissociates from the template DNA would be too short to be accessed by Gre proteins when the nascent RNA is ~10 nt or shorter. However, such a short time should be sufficient to partially block the access of small molecule NTPs into the active site, thereby leading to a short-lived pausing.

On the other hand, the short 6-13-nt transcripts that were produced from the promoters with T-rich ITS were significantly reduced by Gre proteins and by the shorter reaction time (Figure 6 and Supplementary Figure S1), clearly indicating that this signal is involved in abortive synthesis by the moribund complex. The T-rich ITS encoding unstable dA-U hybrid allows energetically favored backtracking to form stable dA-dT duplex [35], likely resulting in abortive poly-U release. The backtracked state originating from the latter mechanism should be much more stable than each of the thermally diffusive backtracked states indicated by the low FEINC.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 August 2020 doi:10.20944/preprints202008.0548.v1

Peer-reviewed version available at Biomolecules 2020, 10, 1299; doi:10.3390/biom10091299

Although further studies are needed to examine whether such in vitro conditions could properly represent the in vivo system, our results suggest that T-rich ITS induces abortive transcription. On the other hand, differently from this latter sequence signal, certain repetitive DNA sequences characterized by the low FEINC induce pausing. Both sequence signals presumably induce backtracking but only in T-rich ITS, the backtracked state becomes more stable than the non-backtracked state.

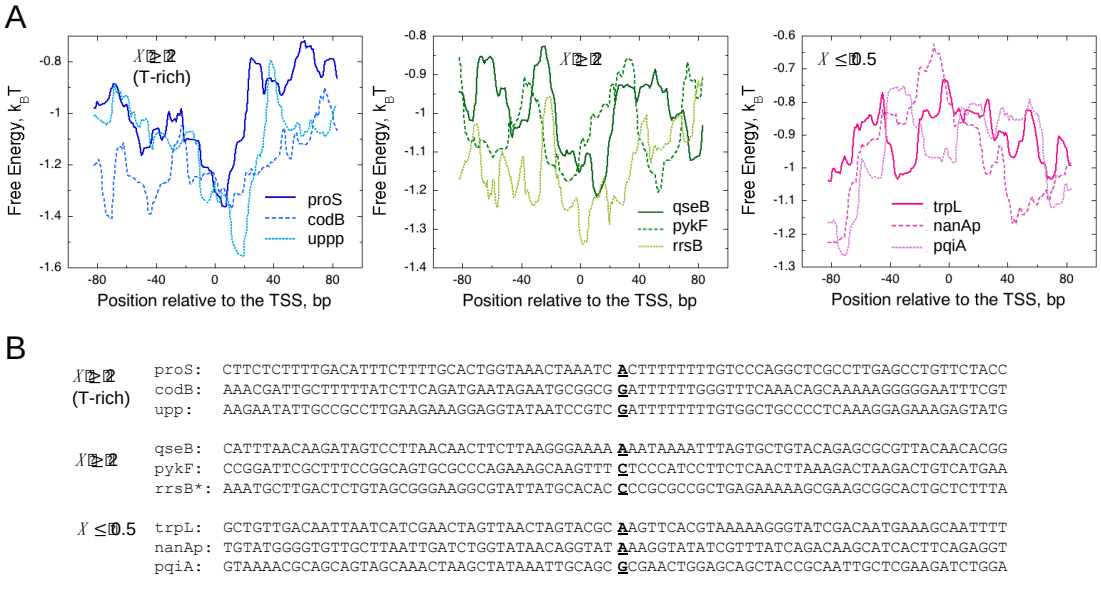


Figure 5. The relation between the FEINC (free energy) landscapes (A) and DNA sequences (B), representing examples of sequences from the thee groups characterized by Table 2. for the RNET-seq reads with T-rich signal, by the ratio $X \geq 2$ without T-rich signal, and by the ratio $X \leq 0.5$, respectively. Each group has three representative promoters in which TSS of the panel B is indicated by bold font with underline. Note that the rrsB promoter, indicated by asterisk, has X=3.81 but was excluded from sequence analyses shown in Figures 3 and 4 due to the redundant property of the rRNA gene.

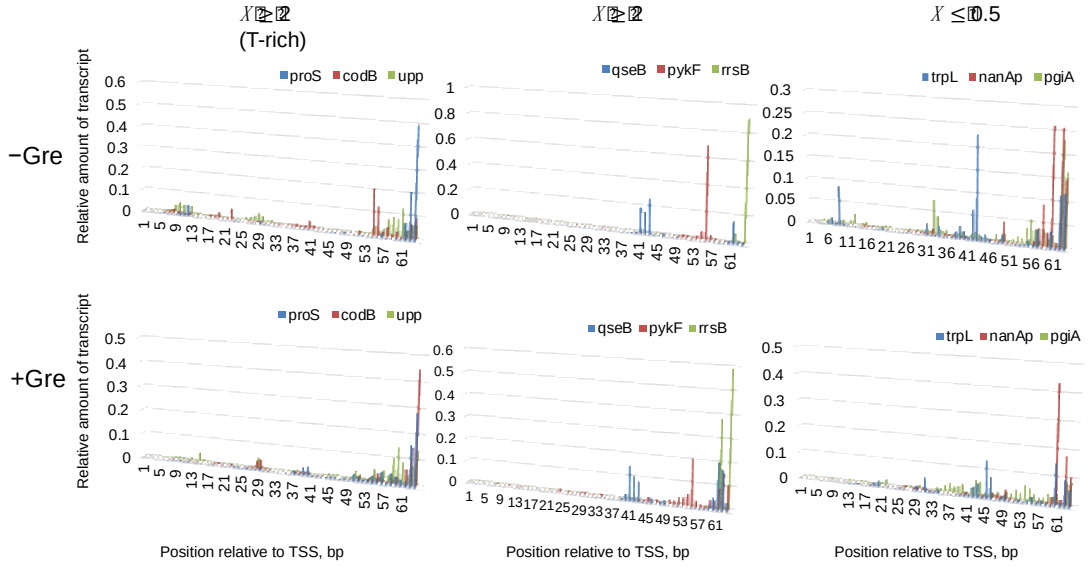
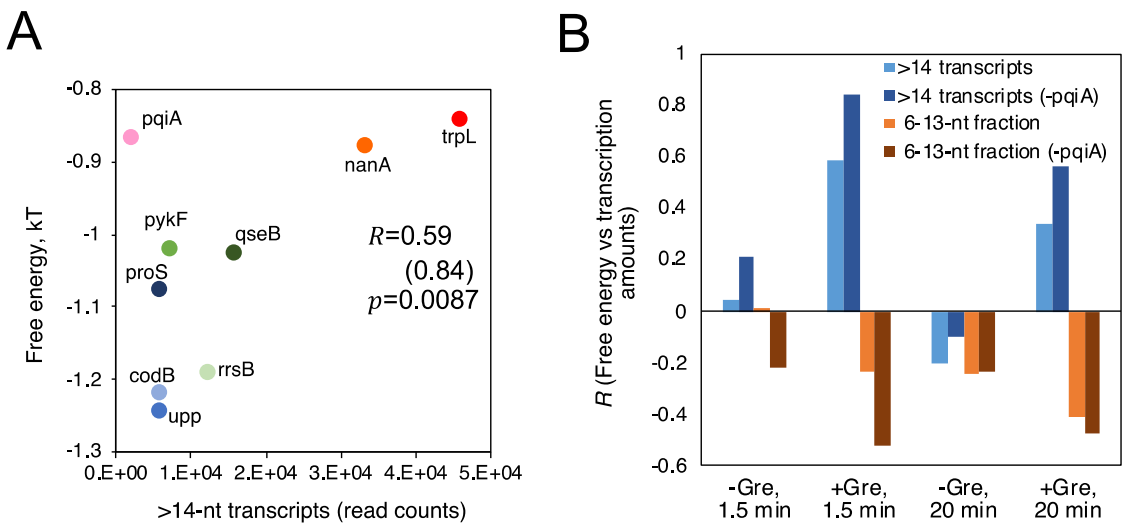


Figure 6. In vitro single-round transcription from 9 different promoters shown in Figure 5B. (A) Entire transcription profiles at 20 min incubation with NTPs are shown in the presence (bottom) or absence (top) of GreAB. See the legend of Figure 5 for the categorization of promoters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 August 2020 doi:10.20944/preprints202008.0548.v1

Peer-reviewed version available at Biomolecules 2020, 10, 1299; doi:10.3390/biom10091299



.

Figure 7. FEINC may predict the productivity of transcription initiation in vitro. (A) Positive correlation of the average FEINC (shown as free energy), which was computed for individual promoters within the interval (-40 bp to +40 bp) around TSS, with the number (read counts) of long >14 nt transcripts. Color code of 9 promoters is same as that of Fig. 5A. Pearson correlation coefficient *R* between the two variables is shown in each graph. *R* values except pqiA are also shown in parentheses as well as the p-value. (B) GreAB and incubation time with NTPs alter the positive and negative correlation trends of the free energy (FEINC) with the number of >14-nt transcripts and the relative fraction of short 6-13 nt transcripts, respectively.

2.3. Sequence and dsDNA-rigidity-related contribution to transcription productivity

In order to investigate the physicochemical properties of dsDNA that determine the productivity of initiation, we performed solution NMR experiment using the 9 promoter DNA of 80 bp shown in Figure 5B. It is generally accepted that imino protons hydrogen-bonded in dsDNA can exchange with water protons only after opening of the base-pairs [36]. Intensities of imino proton resonances detected by NMR experiment indicate how resistant to opening the base-pairs are, and therefore the sum of those intensities through the entire dsDNA can reflect the nonlocal rigidity of dsDNA molecule.

We observed degenerate imino-proton signals derived from individual base pairs in different chemical shift positions (Figure 8A). We assigned those signals to dT-dA base pairs and dG-dC base pairs, respectively, by their dependency on the GC content (Supplementary Figure S2). Since the degenerate signal intensities reflect not only the opening dynamics of base pairs but also the content of base-pairs in dsDNA, we normalized the signal intensities by dividing them by the base-pair content of the dsDNA, focusing only on the rigidity. The normalized signal intensities of dT-dA and dG-dC base pairs globally correlated with each other (Figure 8B), indicating that this value simply reflect nonlocal rigidity of dsDNA. Thus, we define dsDNA rigidity as follows:

$$dsDNA\ ridity = \frac{I_{TA}}{C_{TA}} + \alpha \frac{I_{GC}}{C_{GC}}$$

where I_{TA} and I_{GC} represent the integrated signal intensities of dT-dA and dG-dC base pairs, respectively, and C_{TA} and C_{GC} represent TA and GC contents (%) of the promoter DNA, respectively. A correction coefficient α (~0.033) is obtained by an average value of $\left(\frac{I_{TA}}{C_{TA}}\right)/\left(\frac{I_{GC}}{C_{GC}}\right)$ among 9 promoter DNAs tested, which is used for adjusting the different intensity scale between I_{TA} and I_{GC} due to the different base pair stability. We noticed that dsDNA rigidity tends to negatively correlate with GC content (C_{GC}) through 9 promoter DNA (Figure 8C), suggesting that the maintenance of a certain rigidity in promoter DNA might be physiologically significant especially when the GC content is low.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 August 2020 doi:10.20944/preprints202008.0548.v1

Peer-reviewed version available at Biomolecules 2020, 10, 1299; doi:10.3390/biom10091299

We found that dsDNA rigidity positively correlates with the number (i.e. the read count) of productive >14 nt transcripts (Figure 9A). However, the opposite, negative correlation was observed in the relative fraction of short 6-13 nt transcripts (Figure 9B). The higher the dsDNA rigidity, the higher the productivity of initiation. This trend can be interpreted as indicating that the introduced nonlocal dsDNA rigidity constitutes one of the key reaction coordinates of the productive initiation. We stress that this parameter is entirely sequence-dependent and it is fundamentally determined by nonlocal physicochemical properties of dsDNA. Consistent with the correlative relationship between FEINC and the number of productive (>14 nt) transcripts (Figure 7A), the dsDNA rigidity also positively correlates with the FEINC when pqiA was again excluded as an outlier (Figure 9C). The highest correlation for the number of productive transcripts with the dsDNA rigidity was detected when short (6-13-nt) transcript level was the lowest (by 1.5 min incubation with NTPs in the presence of Gre proteins), while the lowest correlation was obtained when the short transcript level was the highest (by 20 min incubation in the absence of Gre proteins) (Figure 9B). This is similar to the observed correlation between the number of transcripts with FEINC (Figure 7B). These results suggest that the dsDNA rigidity is likely determined to some extent by the presence of repetitive sequence elements.

In summary, our results imply that the pausing-dependent productivity of transcription initiation may be predicted from the two parameters characterizing nonlocal (-40 bp, +40 bp) properties of promoter DNA: (i) an experimentally observable (using NMR) quantity defined here as the dsDNA rigidity and (ii) a calculated measure characterizing repetitive DNA sequence elements defined here as FEINC. These two parameters are presumably connected to each other in terms of the energetics of transcription initiation.

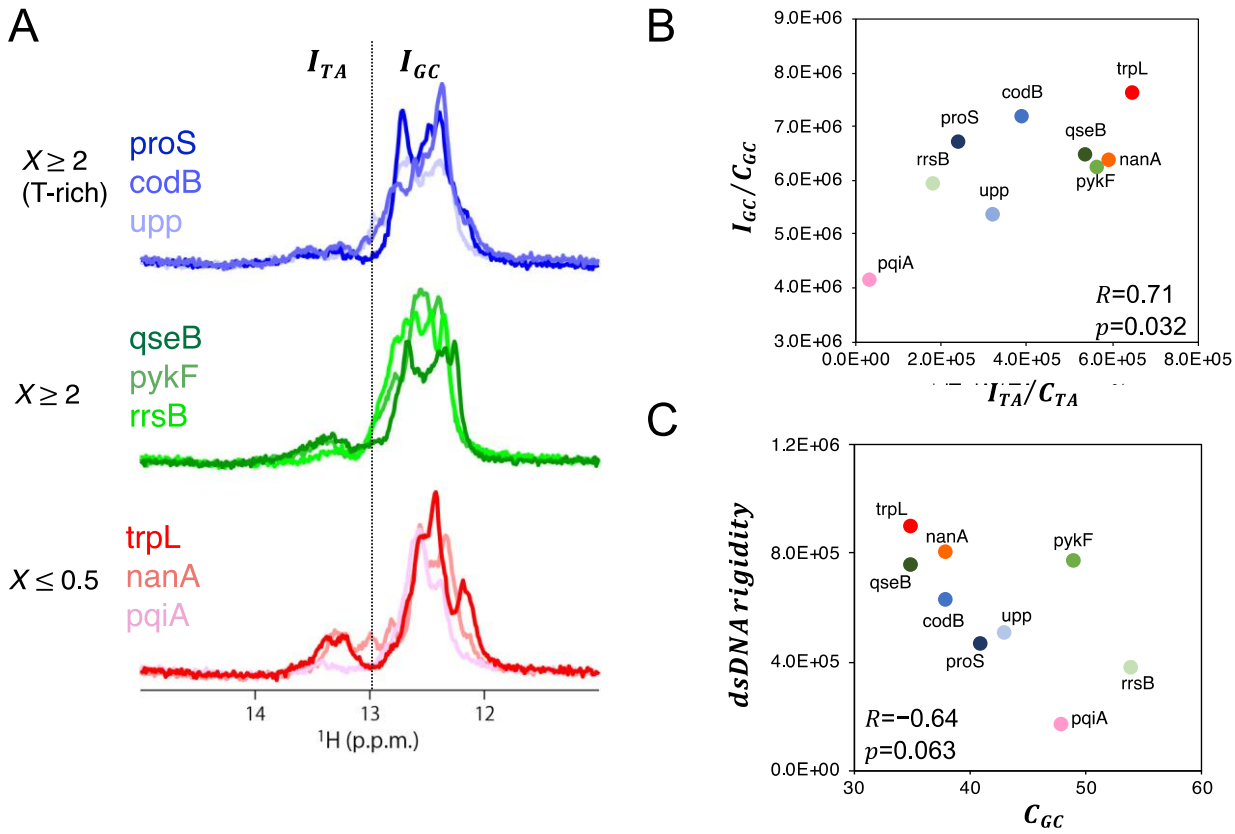


Figure 8. Solution NMR spectroscopy measuring imino proton resonances in 80-bp dsDNA with promoter sequences. (A) Proton NMR spectra of the promoter groups shown in Figure 5B. The dotted line represents the boundary between the imino proton signals that are derived from dT-dA base pair (I_{TA}) and that from dG-dC base pairs (I_{GC}). Larger integrated signal represents overall rigidity between

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 August 2020 doi:10.20944/preprints202008.0548.v1

Peer-reviewed version available at Biomolecules 2020, 10, 1299; doi:10.3390/biom10091299

dsDNA molecules (slow exchange of the imino proton with water proton), while smaller integrated signal represents overall flexibility between dsDNA molecules (rapid exchange of the imino proton with water proton). (B) Global correlation in the base-content-normalized imino proton resonances between dT-dA base pairs (I_{TA}/C_{TA}) and dG-dC base pairs (I_{GC}/C_{GC}) among 9 promoter DNA. (C) dsDNA rigidity tends to negatively correlate with GC contents (C_{GC}) of promoter DNA. Pearson correlation coefficient R and the p-value is shown in each graph of panels B and C.

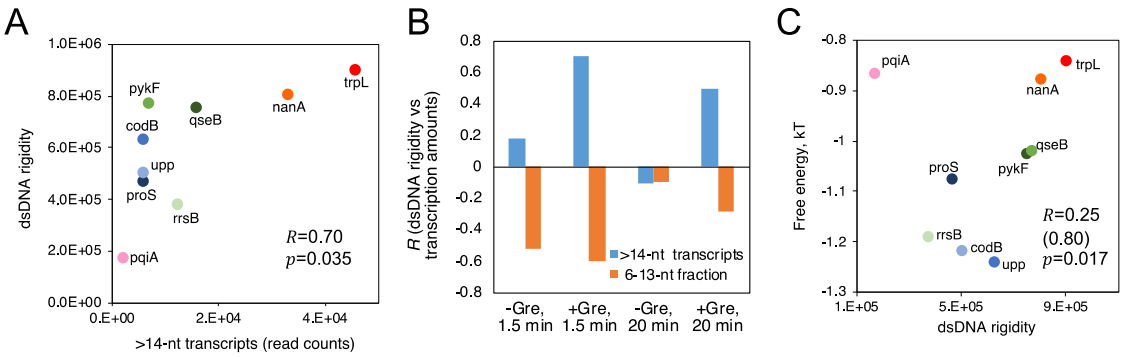


Figure 9. Productivity of transcription depends on nonlocal dsDNA rigidity within promoter region. (A) Correlation analysis between the dsDNA rigidity and the number (read counts) of long >14-nt transcripts. The condition of 1.5 min incubation with NTPs in the presence of Gre proteins was used for the analysis. This condition provided the least abortive transcripts from the 9 promoters on average (see Supplementary Figure S1C). (B) The higher positive and negative correlations of dsDNA rigidity with the long and short transcriptions, respectively, are achieved by the lower production of abortive transcripts. The transcription conditions (±Gre proteins, incubation time with NTPs) are indicated at the bottom of each graph. (C) Correlation analysis between the dsDNA rigidity and the average FEINC (shown as free energy) calculated for individual promoters within the interval (-40 bp to +40 bp) around TSS. In the panel A and C, Pearson correlation coefficient R and the p-value is shown. In the panel C, R value except pqiA is also shown in parentheses.

3. Discussion

In this study, we demonstrated that pausing during initiation is induced by nonlocal interactions between RNAP holoenzyme and promoter DNA with the length of ~80 bp. Our results suggest that such nonlocal interactions are modulated by repetitive DNA sequence elements and are quantified by FEINC, which are also connected to an index of the nonlocal dsDNA rigidity. We identify the nonlocal dsDNA rigidity directly from intensities of imino proton resonances detected by NMR. Therefore, the dsDNA rigidity represents an experimentally identifiable reaction coordinate that depends on physicochemical properties of DNA.

In particular, DNA sequences depleted in repetitive sequence elements (such sequences possess high FEINC, Figure 9C) are characterized by high dsDNA rigidity. In other words, in the higher rigidity region, RNAP-DNA binding turns to favor the consensus mode (sequence-specific binding) more than the nonconsensus mode (sequence-nonspecific binding). Our key finding here is that the high dsDNA rigidity (high FEINC) around σ^{70} promoters appears to play a role in preventing initiation from the pathway decreasing the productivity. More precisely, our results suggest that the high dsDNA rigidity (high FEINC) within the promoter regions can significantly increase the activation energy to the pathways dominated by thermal fluctuations including RNAP sliding on DNA. When dsDNA is less rigid over a wide area including -35 motif, promoter DNA would become more loosely fixed to the σ^{70} subunit of holoenzyme, which allows positional fluctuations (i.e. sliding) between RNAP and DNA. These fluctuations interfere with the σ^{70}-specific promoter recognition during the initial binary complex formation, and also induce backtracking in the following stage of the ternary complex. The diffusive backtracking impedes forward translocation that is necessary for the next NTP addition to the nascent 3′ RNA end during progressing transcription. This results in pausing. Indeed, in the promoters with a smaller dsDNA rigidity, we observed decreased amounts

of promoter-specific transcriptions and increased lifetime of pausing during promoter escape (Figure 9 A and B).

We also identified T-rich ITS as a sequence motif weakly conserved for increasing abortive transcription genome-wide. Such T-rich ITS encodes dA-U hybrid that is less stable than its dsDNA form dA-dT duplex [35], thereby allowing abortive release of oligomeric U by backtracking. After releasing of oligomeric U, a more stable dA-dT DNA duplex can form. In this sense, the backtracking induced by the signal is energetically favored and is different from that induced by thermal fluctuations on repetitive sequences. We verified that the T-rich ITS causes abortive initiation using the in vitro transcription assay.

3.1. FEINC predicts conformational heterogeneity of moribund complexes

One of the key advantages of our statistical approach using the RNET-seq data is that it allows us to analyze the conformational heterogeneity of the ternary complexes in vivo, depending on the length of the nascent RNA retained. Focusing on the abortive/pausing-enriched promoter group (Figure 4, $X \geq 2$), we found an opposite trend in the landscape of the average FEINC for the complexes retaining 7-nt RNA, as compared to the complexes retaining 8-nt RNA (Figure 10A).

In particular, the free energy (i.e. FEINC) of the 7-nt RNA retained complex was higher compared to the corresponding free energy of the ≥8-nt RNA retained complex, within the range of DNA sequence forming the DNA bubble and the RNA-DNA hybrid (~-11 to +7). However conversely, the free energy of the 7-nt RNA retained complex was lower through the upstream and downstream dsDNA flanking the bubble and the hybrid (Figure 10 A and B). Such dsDNA-RNAP interactions include (i) σ region 4 and -35 DNA motif, (ii) the C-terminal domain of the α subunit (α-CTD) and the upstream DNA (UP-element), and (iii) so-called jaw domain of the β′ subunit and the downstream DNA (Figure 10 A and B) [37]. The increased nonconsensus binding mode through the broad range of dsDNA interactions with the protein surface may loosen the strong interaction stemming from σ region 3.2 loop and phosphates of the 5′ nascent transcript by increasing the sliding between the protein and dsDNA [20]. Indeed, single-molecule studies of different research group have identified a long-lived pausing that likely stems from the σ-5′ RNA interaction on lacCONS promoter when the growing RNA reaches 7-nt [8,9,20]. Interestingly, the pausing observed was involved in 1-bp backtracking [20]. When the RNA was extended to ≥8 nt, the lower free energy region was more localized to the hybrid and the bubble if compared to that of the 7-nt RNA retained complex, and was shifted to the upstream when the RNA was further extended to 9 nt and 10 nt.

Overall, the RNA-length-dependent difference in the FEINC landscapes suggests that the functional significance of the nonconsensus mode in the RNAP-dsDNA-hybrid binding varies depending on the nascent RNA length. Unlike the nonconsensus mode, no striking difference was observed in the -10/-35 consensus motifs between the ternary complexes of the abortive/pausing-enriched group carrying 7-nt and 8-nt RNAs (Figure 3 A and B). This result suggests that the nonconsensus mode predicted by the FEINC landscape rather than the conventional consensus motifs mainly contributes to determine the RNA-length-dependent conformational heterogeneity.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 August 2020 doi:10.20944/preprints202008.0548.v1

Peer-reviewed version available at Biomolecules 2020, 10, 1299; doi:10.3390/biom10091299

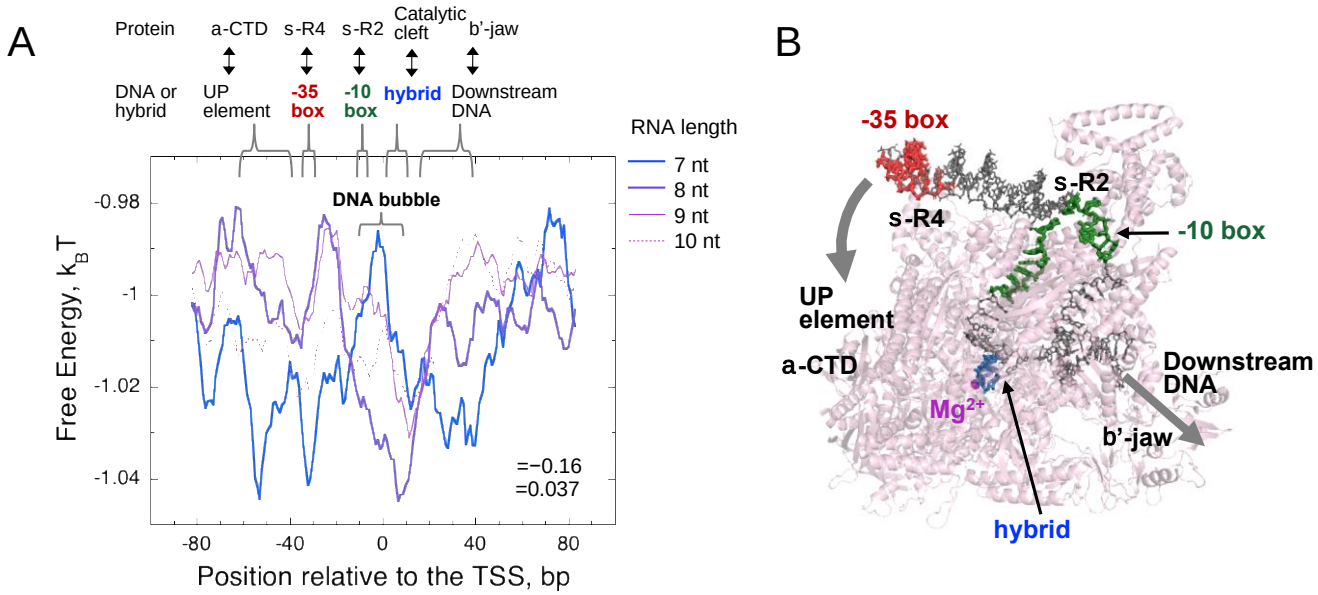


Figure 10. FEINC landscape is altered by the nascent RNA length. (A) An opposite trend of the FEINC landscapes is observed between the ternary complexes retaining 7-nt and 8-nt (9 nt and 10 nt) RNA, in the abortive/pausing-enriched group ($X \geq 2$). Pearson correlation coefficient R and the p-value in the comparison of the entire free energy indices between the complexes having 7-nt RNA and 8-nt RNA are shown in the graph. (B) X-ray crystal structure of the *E. coli* initiation complex with 4-bp RNA-DNA hybrid (PDB ID: 4YLN) [38]. Key interactions between holoenzyme and dsDNA/RNA-DNA hybrid that are described in the panel A are also shown within the structure. DNA and holoenzyme molecules are shown by gray and pink colors, respectively.

4. Conclusion

Our statistical analysis of transcription initiation using the concept of the nonconsensus mode of protein-DNA binding suggests that the fate of the nascent transcript on σ^{70} promoters is determined, at least partially, by repetitive DNA sequence elements around TSS. Repetitive sequence elements can increase the number of possible conformational states of the ternary complex including backtracking. We suggest that certain types of repetitive elements can also decrease the productivity of initiation by lowering dsDNA rigidity. Therefore, we argue that the definition of functional promoter sequences should be reconsidered, including quantitative measures accounting for the effect of repetitive sequence elements and nonlocal dsDNA rigidity.

Finally, we demonstrate here that concepts of statistical mechanics provide a firm theoretical framework for handling high-throughput sequencing data containing the information on microscopic heterogeneity of a protein-DNA-RNA ternary complex [29]. In this study, we demonstrate that such a dynamic property of macromolecules in aqueous solution can be directly accessed by solution NMR experiment. In future, such a combination of NMR with the statistical approach should further uncover interesting but unexplained phenomena that still remain in the field of transcription.

5. Materials and Methods.

5.1. In vitro transcription.

E. coli RNAP holoenzyme, GreA and GreB proteins were purified as described previously [39,40]. NTPs and oligonucleotides were purchased from GE Healthcare and Fasmac, respectively. The liner DNA templates from −77 to +63 when TSS is +1, each of which contains one of 9 promoters

shown in Figure 5B, was prepared by PCR using oligonucleotides (see Table S1 for the full sequences). These DNA templates were purified by PAGE.

All reactions were performed in transcription buffer (TB; 20 mM Tris–HCl, pH 7.6, 5 mM $MgCl_2$, 1 mM 2-mercaptoethanol, 0.1 M KCl) at room temperature. The holoenzyme (200 nM) and DNA template (10 nM each of 9 promoter DNA) were preincubated for 10 min in TB. Where present, GreA and GreB proteins were added to the holoenzyme at final concentration of 7 μM and 4 μM, respectively. Reaction was started by adding 100 μM NTPs at final concentration. Heparin (250 μg/ml) was added together with the substrates to eliminate enzyme turnover, which assures single-round reaction. After incubation for 1.5 min or 20 min, reaction was stopped by adding phenol/chloroform/isoamyl alcohol (25 : 24 : 1). The experiments with presented results in this study were repeated twice and the represented ones are shown.

RNA transcripts were analyzed by Illumina sequencing. Briefly, cDNA libraries were constructed according to [21]. Quantification of the cDNAs was performed by RT-PCR. Illumina sequencing was performed with MiniSeq High Output Kit (75 Cycles). A typical output of the sequencing was ~3 x 10^6 reads per sample. The number of 3' RNAs that were mapped and aligned to DNA template was counted as described previously [21].

5.2. Solution NMR.

DNA oligonucleotides that were purified by reverse phase cartridge were purchased from Fasmac. Each double-stranded DNA (dsDNA) of 50 μM was generated in a 250 μL solution consisting of 90% H_2O/10% D_2O solvent with 20 mM Tris-D11 (pH 7.6 at 25°C), 5 mM $MgCl_2$, and 50 mM KCl, which was transferred into a 5-mm microtube (Shigemi, Tokyo). NMR experiments were performed on an Avance 700 spectrometer (Bruker, Billerica, MA) equipped with a 5-mm TXI triple resonance probe at 15, 25, and 35°C. Proton one-dimensional spectra were recorded with a 22 ppm spectral width, centered at 4.7 ppm, using the WATERGATE building block for solvent suppression. Sodium 3-(trimethylsilyl)-1-propanesulfonate was used as an external chemical shift standard. Free induction decays (FIDs) were acquired for 133 ms with 2,048 digital points. The FIDs were accumulated 4096 times with interscan delays of 4.0 (15°C) and 2.5 s (25 and 35°C). Raw FIDs were multiplied by a cosine window function and Fourier transformed to frequency domain data, followed by phase correction and baseline correction. Signals in the chemical shift range of 11.6-14.2 ppm were integrated, which we used as the imino proton signal intensity.

Acknowledgements: We thank Mikhail Kashlev and Lucyna Lubkowska for helpful discussions and for E. coli RNAP and Gre proteins. We also thank Koh Takeuchi for critical reading of the manuscript. This work was financially supported by JSPS KAKENHI Grant 18K18731 and 20H03298 (to M.I.).

References

1. Saecker, R.M.; Record, M.T.; Dehaseth, P.L. Mechanism of bacterial transcription initiation: RNA polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. *J Mol Biol* **2011**, *412*, 754-771, doi:10.1016/j.jmb.2011.01.018.
2. Browning, D.F.; Busby, S.J. The regulation of bacterial transcription initiation. *Nat Rev Microbiol* **2004**, *2*, 57-65, doi:10.1038/nrmicro787.
3. Einav, T.; Phillips, R. How the avidity of polymerase binding to the -35/-10 promoter sites affects gene expression. *Proc Natl Acad Sci U S A* **2019**, *116*, 13340-13345, doi:10.1073/pnas.1905615116.
4. Johnston, D.E.; McClure, W.R. *Abortive initiation of in vitro RNA synthesis on bacteriophage DNA.*; Cold Spring Harbor Laboratory Press,: NY, 1976; pp. 413-428.
5. Carpousis, A.J.; Gralla, J.D. Cycling of ribonucleic acid polymerase to produce oligonucleotides during initiation in vitro at the lac UV5 promoter. *Biochemistry* **1980**, *19*, 3245-3253, doi:10.1021/bi00555a023.
6. Kubori, T.; Shimamoto, N. A branched pathway in the early stage of transcription by Escherichia coli RNA polymerase. *J. Mol. Biol.* **1996**, *256*, 449-457.
7. Sen, R.; Nagai, H.; Shimamoto, N. Polymerase arrest at the lambdaP(R) promoter during transcription initiation. *J Biol Chem* **2000**, *275*, 10899-10904, doi:10.1074/jbc.275.15.10899.

8. Lerner, E.; Chung, S.; Allen, B.L.; Wang, S.; Lee, J.; Lu, S.W.; Grimaud, L.W.; Ingargiola, A.; Michalet, X.; Alhadid, Y., et al. Backtracked and paused transcription initiation intermediate of Escherichia coli RNA polymerase. *Proc Natl Acad Sci U S A* **2016**, *113*, E6562-E6571, doi:10.1073/pnas.1605038113.
9. Duchi, D.; Bauer, D.L.; Fernandez, L.; Evans, G.; Robb, N.; Hwang, L.C.; Gryte, K.; Tomescu, A.; Zawadzki, P.; Morichaud, Z., et al. RNA Polymerase Pausing during Initial Transcription. *Mol Cell* **2016**, *63*, 939-950, doi:10.1016/j.molcel.2016.08.011.
10. Dulin, D.; Bauer, D.L.V.; Malinen, A.M.; Bakermans, J.J.W.; Kaller, M.; Morichaud, Z.; Petushkov, I.; Depken, M.; Brodolin, K.; Kulbachinskiy, A., et al. Pausing controls branching between productive and non-productive pathways during initial transcription in bacteria. *Nat Commun* **2018**, *9*, 1478, doi:10.1038/s41467-018-03902-9.
11. Samanta, S.; Martin, C.T. Insights into the mechanism of initial transcription in Escherichia coli RNA polymerase. *J Biol Chem* **2013**, *288*, 31993-32003, doi:10.1074/jbc.M113.497669.
12. Shimamoto, N. Nanobiology of RNA polymerase: biological consequence of inhomogeneity in reactant. *Chem Rev* **2013**, *113*, 8400-8422, doi:10.1021/cr400006b.
13. Susa, M.; Kubori, T.; Shimamoto, N. A pathway branching in transcription initiation in Escherichia coli. *Mol. Microbiol.* **2006**, *59*, 1807-1817.
14. Henderson, K.L.; Felth, L.C.; Molzahn, C.M.; Shkel, I.; Wang, S.; Chhabra, M.; Ruff, E.F.; Bieter, L.; Kraft, J.E.; Record, M.T. Mechanism of transcription initiation and promoter escape by. *Proc Natl Acad Sci U S A* **2017**, *114*, E3032-E3040, doi:10.1073/pnas.1618675114.
15. Imashimizu, M.; Tanaka, K.; Shimamoto, N. Comparative Study of Cyanobacterial and E. coli RNA Polymerases: Misincorporation, Abortive Transcription, and Dependence on Divalent Cations. *Genet Res Int* **2011**, *2011*, 572689, doi:10.4061/2011/572689.
16. Sen, R.; Nagai, H.; Shimamoto, N. Conformational switching of Escherichia coli RNA polymerase-promoter binary complex is facilitated by elongation factor GreA and GreB. *Genes Cells* **2001**, *6*, 389-401.
17. Borukhov, S.; Sagitov, V.; Goldfarb, A. Transcript cleavage factors from E. coli. *Cell* **1993**, *72*, 459-466, doi:10.1016/0092-8674(93)90121-6.
18. Skancke, J.; Bar, N.; Kuiper, M.; Hsu, L.M. Sequence-Dependent Promoter Escape Efficiency Is Strongly Influenced by Bias for the Pretranslocated State during Initial Transcription. *Biochemistry* **2015**, *54*, 4267-4275, doi:10.1021/acs.biochem.5b00272.
19. Heyduk, E.; Heyduk, T. DNA template sequence control of bacterial RNA polymerase escape from the promoter. *Nucleic Acids Res* **2018**, *46*, 4469-4486, doi:10.1093/nar/gky172.
20. Lerner, E.; Ingargiola, A.; Lee, J.J.; Borukhov, S.; Michalet, X.; Weiss, S. Different types of pausing modes during transcription initiation. *Transcription* **2017**, *8*, 242-253, doi:10.1080/21541264.2017.1308853.
21. Imashimizu, M.; Takahashi, H.; Oshima, T.; McIntosh, C.; Bubunenko, M.; Court, D.L.; Kashlev, M. Visualizing translocation dynamics and nascent transcript errors in paused RNA polymerases in vivo. *Genome Biol* **2015**, *16*, 98, doi:10.1186/s13059-015-0666-5.
22. Imashimizu, M.; Afek, A.; Takahashi, H.; Lubkowska, L.; Lukatsky, D.B. Control of transcriptional pausing by biased thermal fluctuations on repetitive genomic sequences. *Proceedings of the National Academy of Sciences of the United States of America* **2016**, *113*, E7409-E7417, doi:10.1073/pnas.1607760113.
23. Santos-Zavaleta, A.; Salgado, H.; Gama-Castro, S.; Sánchez-Pérez, M.; Gómez-Romero, L.; Ledezma-Tejeida, D.; García-Sotelo, J.S.; Alquicira-Hernández, K.; Muñiz-Rascado, L.J.; Peña-Loredo, P., et al. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Res* **2019**, *47*, D212-D220, doi:10.1093/nar/gky1077.
24. Stepanova, E.; Lee, J.; Ozerova, M.; Semenova, E.; Datsenko, K.; Wanner, B.L.; Severinov, K.; Borukhov, S. Analysis of promoter targets for Escherichia coli transcription elongation factor GreA in vivo and in vitro. *J Bacteriol* **2007**, *189*, 8772-8785, doi:10.1128/JB.00911-07.
25. Schneider, T.D.; Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **1990**, *18*, 6097-6100, doi:10.1093/nar/18.20.6097.
26. Schurr, J.M. The one-dimensional diffusion coefficient of proteins absorbed on DNA. Hydrodynamic considerations. *Biophys Chem* **1979**, *9*, 413-414.
27. Sela, I.; Lukatsky, D.B. DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. *Biophys J* **2011**, *101*, 160-166, doi:10.1016/j.bpj.2011.04.037.
28. Afek, A.; Schipper, J.L.; Horton, J.; Gordân, R.; Lukatsky, D.B. Protein-DNA binding in the absence of specific base-pair recognition. *Proc Natl Acad Sci U S A* **2014**, *111*, 17140-17145, doi:10.1073/pnas.1410569111.

29. Imashimizu, M.; Lukatsky, D.B. Transcription pausing: biological significance of thermal fluctuations biased by repetitive genomic sequences. *Transcription* **2018**, *9*, 196-203, doi:10.1080/21541264.2017.1393492.
30. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **2002**, *12*, 656-664, doi:10.1101/gr.229202.
31. Imashimizu, M.; Kireeva, M.L.; Lubkowska, L.; Gotte, D.; Parks, A.R.; Strathern, J.N.; Kashlev, M. Intrinsic Translocation Barrier as an Initial Step in Pausing by RNA Polymerase II. *Journal of molecular biology* **2013**, *425*, 697-712, doi:10.1016/j.jmb.2012.12.002.
32. Imashimizu, M.; Oshima, T.; Lubkowska, L.; Kashlev, M. Direct assessment of transcription fidelity by high-resolution RNA sequencing. *Nucleic acids research* **2013**, *41*, 9090-9104, doi:10.1093/nar/gkt698.
33. Komissarova, N.; Becker, J.; Solter, S.; Kireeva, M.; Kashlev, M. Shortening of RNA:DNA hybrid in the elongation complex of RNA polymerase is a prerequisite for transcription termination. *Mol Cell* **2002**, *10*, 1151-1162.
34. Shimada, T.; Yamazaki, Y.; Tanaka, K.; Ishihama, A. The whole set of constitutive promoters recognized by RNA polymerase RpoD holoenzyme of Escherichia coli. *PLoS One* **2014**, *9*, e90447, doi:10.1371/journal.pone.0090447.
35. Huang, Y.; Weng, X.; Russu, I.M. Structural energetics of the adenine tract from an intrinsic transcription terminator. *J Mol Biol* **2010**, *397*, 677-688, doi:10.1016/j.jmb.2010.01.068.
36. Gueron, M.; Kochoyan, M.; Leroy, J.L. A single mode of DNA base-pair opening drives imino proton exchange. *Nature* **1987**, *328*, 89-92, doi:10.1038/328089a0.
37. Ruff, E.F.; Record, M.T.; Artsimovitch, I. Initial events in bacterial transcription initiation. *Biomolecules* **2015**, *5*, 1035-1062, doi:10.3390/biom5021035.
38. Zuo, Y.; Steitz, T.A. Crystal structures of the E. coli transcription initiation complexes with a complete bubble. *Mol Cell* **2015**, *58*, 534-540, doi:10.1016/j.molcel.2015.03.010.
39. Imashimizu, M.; Kireeva, M.L.; Lubkowska, L.; Kashlev, M.; Shimamoto, N. The Role of Pyrophosphorolysis in the Initiation-to-Elongation Transition by E. coli RNA Polymerase. *J Mol Biol* **2019**, doi:10.1016/j.jmb.2019.04.020.
40. Imashimizu, M.; Oshima, T.; Takahashi, H.; Lubkowska, L.; Kashlev, M. Direct Assessment of Transcription Fidelity by RNA Sequencing. *Biophysical Journal* **2014**, *106*, 486A-486A, doi:10.1016/j.bpj.2013.11.4468.