Physical Biology

PAPER

Density- and elongation speed-dependent error correction in RNA polymerization

RECEIVED
8 March 2021

REVISED
3 November 2021

ACCEPTED FOR PUBLICATION
22 December 2021

PUBLISHED
25 January 2022

Xinzhe Zuo[1] and Tom Chou[1,2,*]

[1] Department of Mathematics, UCLA, Los Angeles, CA 90095-1555, United States of America
[2] Department of Computational Medicine, UCLA, Los Angeles, CA 90095-1766, United States of America
* Author to whom any correspondence should be addressed.

E-mail: tomchou@ucla.edu

Keywords: RNA polymerase, backtracking, error correction, first passage times, stochastic model

Abstract

Backtracking of RNA polymerase (RNAP) is an important pausing mechanism during DNA transcription that is part of the error correction process that enhances transcription fidelity. We model the backtracking mechanism of RNAP, which usually happens when the polymerase tries to incorporate a noncognate or 'mismatched' nucleotide triphosphate. Previous models have made simplifying assumptions such as neglecting the trailing polymerase behind the backtracking polymerase or assuming that the trailing polymerase is stationary. We derive exact analytic solutions of a stochastic model that includes locally interacting RNAPs by explicitly showing how a trailing RNAP influences the probability that an error is corrected or incorporated by the leading backtracking RNAP. We also provide two related methods for computing the mean times for error correction and incorporation given an initial local RNAP configuration. Using these results, we propose an effective interacting-RNAP lattice that can be readily simulated.

1. Introduction

Transcription is the first step of DNA-based gene expression. During the transcription process, an RNA polymerase (RNAP) enzyme binds and separates the ds-DNA, forming a transcription bubble or elongation complex at a promoter site. As the RNAP and its transcription elongation complex (TEC) move along the DNA, additional RNAPs can initiate new complexes at the empty promoter site. Each RNAP processes along the DNA up to the termination site, adding nucleotides to the 3′ end of the newly formed RNA transcript along the way. The TEC forms an exclusionary zone similar to that seen in a chain of ribosomes translating mRNA during protein production. Thus, it would be natural to apply stochastic models such as the totally asymmetric exclusion process (TASEP) originally developed for studying mRNA translation [1–6] to the DNA transcription process.

While DNA replication by DNA polymerase results in an error rate of 10^{-8} to 10^{-10} per base pair [7–9], RNAP has a much higher error rate of 10^{-4} to 10^{-6} per base pair [10–14]. Since some RNAs are present at a level of less than one molecule per cell in microbes [15] and in embryonic stem cells [16], a gene may be represented by a single mutated RNA transcript. Therefore, error-free transcription plays an important role in faithful gene expression.

RNAPs are sometimes interrupted by pauses [17, 18]. Krummel observed irregular DNA footprints suggesting that RNAP shrinks and expands during the elongation process [19]. From this observation, an 'inchworming' model for the elongation of RNAP was developed [20]. However, later experiments suggested that the inchworming phenomenon was actually the RNAP complex traveling back and forth along the DNA template [21–23]. Now known as RNAP backtracking, this important pausing mechanism aids proofreading and fidelity of the transcription process. Backtracking strongly depends on the stability of the RNA/DNA hybrid in the TEC; the weaker the hybrid, the higher the probability for backtracking [21]. Hence, when a wrong nucleotide triphosphate (NTP) is added to the transcript, the 3′ end of the RNAP is frayed, which induces backtracking.

During backtracking, the 3′ end of the RNA disengages with the RNAP catalytic site, rendering the RNAP complex inactive but stable [21, 24]. Figure 1 depicts a chain of RNAPs, their associated nascent RNA transcripts, and one erroneous nucleotide (red asterisk). We assume that once a wrong nucleotide is added to the catalytic site, the RNAP enters a backtracking state during which it can move backwards relative to both the DNA and the RNA transcript without depolymerizing the transcript. As a result, the 3′ end of the RNA transcript now extrudes out of the RNAP. There are two competing processes for the RNAP to exit the backtracking state, as depicted in the lower insets of figure 1. In one, the RNAP can perform a random walk on the DNA template until realignment occurs [25–27] and the erroneous nucleotide is incorporated into the transcript. In the other, a segment of transcript associated with backtracking RNAP can be cleaved so that a new RNA 3′ end which aligns with the active site is created [28–30]. In eukaryotic and prokaryotic cells, transcript cleavages are enhanced by cleavage factors TFIIS [31, 32] and GreA/GreB [33], respectively. Cleavage of the mismatched nucleotide before incorporation allows the transcript under construction to be corrected [18].

Previous theoretical studies have analyzed the kinetics of backtracking of a single RNAP as elongation occurs, giving rise to non-Poissonian pause times [34] and bursty mRNA production [35, 36]. Roldán *et al* [37] examined the mean depth and time of backtracking RNAPs under both discrete and continuous semi-infinite chains, while Sahoo and Klump [38] studied the accuracy of the transcription in the context of a single RNAP. Both studies assumed that the trailing RNAP is stationary. However, when the leading RNAP is in a backtracking state, the trailing RNAP is not stationary and would most likely be in an actively processing state that closes the gap between it and the leading RNAP [39–42]. The closing of a trailing RNAP might be described in terms of a 'pushing' mechanism which is one finding of our subsequent analysis. Note, the 'pushing' mechanism implied in our site exclusion model is entirely entropic and distinct from the actual mechanical pushing suggested in [43].

Another class of models developed to explain bursty transcription incorporates positive and negative DNA supercoiling downstream and upstream of an RNAP complex, respectively [36, 44, 45]. Effects of supercoiling for multiple RNAPs have also been studied in [46, 47], which shows that supercoiling could increase the variance of the number of RNAPs on a DNA template during transcription. Supercoiling arises when both RNAPs and their nearby DNA are prevented from rotating/twisting. To generate supercoiling, RNAPs must move in part via a power stroke and carry an appreciably rotational drag [47], while DNA must be constrained and any accumulated twist is not dissipated by topoisomerases and gyrases.

If the RNAP moves forward as a result of a Brownian ratchet, we would expect a low Stokes efficiency [48] and very little correlated mechanical twisting required to appreciably supercoil DNA. Only if the RNAP processes predominantly through a power stroke mechanism could supercoiling build up to hinder further power stroke-induced elongation. However, if the supercoiling can relax or dissipate between each power stroke, it would not appreciable hinder subsequent elongation. This would occur locally during the time an RNAP is in the 'diffusive' backtracking state under which no energy is being used. Moreover, supercoiling has typically been observed under isolated *in vitro* conditions in which the RNAP is anchored and the DNA is constrained, although some *in vivo* observations have also recently been made [49]. In our analysis, we will neglect explicit mechanical effects arising from supercoiling by assuming that viscous drag directly hinders the RNAP and/or, that any residual DNA supercoiling is dissipated by DNA rearrangement and/or gyrase/topoisomerase activity such that appreciable supercoiling is not reached, allowing us to assume that the RNAP elongation rates, backtracking diffusivity, and other parameters are constant (translationally invariant).

Although interacting lattice models derived from variations of a TASEP have been applied to multi-RNAP mediated transcription [35, 50, 51], they do not include error correction through the pausing and backtracking mechanism. Assuming the discussed assumptions, we derive and solve a local discrete stochastic model that incorporates a trailing RNAP that closes in on the leading one, allowing us to better understand how interactions between neighboring RNAPs influence the probabilities and timescales of error correction. Our exact two-particle interaction model can then be systematically incorporated into lattice models to study aspects of collective, interacting-RNAP transcription.

2. Local stochastic model

Consider RNAPs with effective size ℓ (which includes the associated TEC) that normally process along the gene at rate p as shown in figure 2. For clarity, the RNA transcripts emanating from the RNAPs are not shown. We now focus on two adjacent RNAPs: a leading one that has just recruited a wrong nucleotide (at the position marked by the red asterisk) and a trailing one just upstream (behind) of the leading RNAP.

The nucleotide mismatch promotes transition of the leading RNAP into the backtracking state [18, 52]. Immediately after a noncognate nucleotide is presented, the RNAP can continue on and incorporate it with probability P^* or, enter a backtracking state with probability $1 - P^*$. In the backtracking

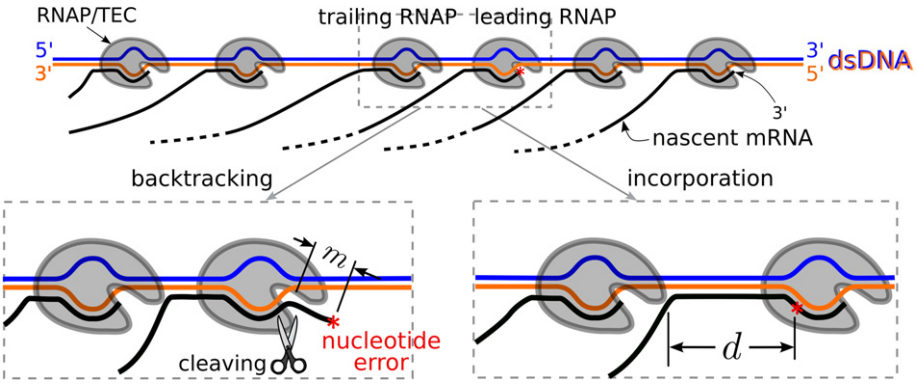

Figure 1. Schematic of interacting RNAPs. A chain of transcribing RNAPs processing along DNA. Each RNAP forms a local TEC through which one strand is copied to RNA. If an erroneous nucleotide (red asterisk) is recruited, the RNAP can either incorporate the error (lower right inset) or backtrack along the DNA (lower left inset). During backtracking, the RNAP loosens its grip and the active site reverses along the transcribed RNA in a random-walk-like fashion, allowing it to cleave off the misincorporated nucleotide (lower left inset). In the backtracking state, cleaving near the 3′ end of the mRNA can excise the error. The distance between the misincorporated nucleotide site on the DNA and the leading edge of the trailing RNAP is denoted d (lower right inset), while the accessible distance between the back edge of the leading RNAP and the leading edge of the trailing RNAP is denoted n. If the leading RNAP backtracks and exposes the nucleotide error at the 3′ end, the extruded distance is denoted m (lower left inset). Our stochastic model is fully described by the discrete integers (m, n).

state, the leading RNAP can undergo a symmetric random walk with hopping rate q across the accessible sites between the trailing RNAP and the position of the mismatched nucleotide (red asterisk). During this backtracking, if the leading RNAP is abutted at the misincorporated nucleotide site, it can hop across it with rate $k_{\rm inc}$ and permanently incorporate the wrong nucleotide into the nascent mRNA. Thus, there are two ways of incorporating a noncognate nucleotide: immediate incorporation with probability P^* or incorporation that occurs after entering into a backtracking state.

In the backtracking state, another outcome is possible. During diffusive motion of the leading RNAP, the end fragment of the mRNA transcript can also be cleaved with rate $k_{\rm c}$, removing the erroneous NTP and rescuing the leading RNAP from the backtracking state as it resumes elongation. At the same time, the trailing RNAP is still moving forward with rate p if it is unblocked by the leading RNAP.

To construct our model, we will condition our analysis on the RNAP entering the backtracking state. As depicted in figure 2 (top), we define m to be the distance between the leading RNAP and the realignment position associated with the noncognate nucleotide. Let n be the number of accessible sites available to the backtracking RNAP (the distance d between the error site and the trailing RNAP, minus the number of sites ℓ occluded by the leading RNAP bubble). Thus, n can only decrease when the trailing RNAP advances. The state variables (m, n) are the effective distances between the erroneous nucleotide site and the leading and trailing RNAPs, respectively. Starting from an initial condition in which a wrong nucleotide has just been added and the leading RNAP has just entered into a backtracking state $(m = 0)$, the evolution of the system can be described by the state diagram in figure 2 (bottom). For the interior points, $m \geqslant 1$ and $n > m$,

$$\frac{{\rm d}P_n(m,t)}{{\rm d}t} = -(k_{\rm c} + 2q + p)P_n(m,t) + pP_{n+1}(m,t) + qP_n(m+1,t) + qP_n(m-1,t). \quad (1)$$

For the boundary states $m = 0$, $n \geqslant 1$,

$$\frac{{\rm d}P_n(0,t)}{{\rm d}t} = -(k_{\rm inc} + p + q)P_n(0,t) + qP_n(1,t) + pP_{n+1}(0,t), \quad (2)$$

while the probabilities of the edge states $m = n$ obey

$$\frac{{\rm d}P_n(n,t)}{{\rm d}t} = pP_{n+1}(n,t) + qP_n(n-1,t) - (k_{\rm c} + q)P_n(n,t), \quad (3)$$

and that of the corner point obeys

$$\frac{{\rm d}P_0(0,t)}{{\rm d}t} = pP_1(0,t) - k_{\rm inc}P_0(0,t). \quad (4)$$

The initial condition, defined at the instant a wrong nucleotide is added is $P_n(m, t = 0) = 1$ $(m, 0)1(n, N)$. Solution of equations (1)–(3) yields the probability the system is in state (m, n) at time t.

2.1. Iterative solution for $n = N$

First, consider an initial fixed distance $n = N$ between the trailing RNAP and the site of misincorporation (see top panel, figure 2). Since the trailing RNAP can move only forward, the $n = N$ chain provides a source of probability flux into the $n = N - 1$ chain. From the probabilities distributed across the $n = N$ chain, we can calculate the time-dependent probability fluxes

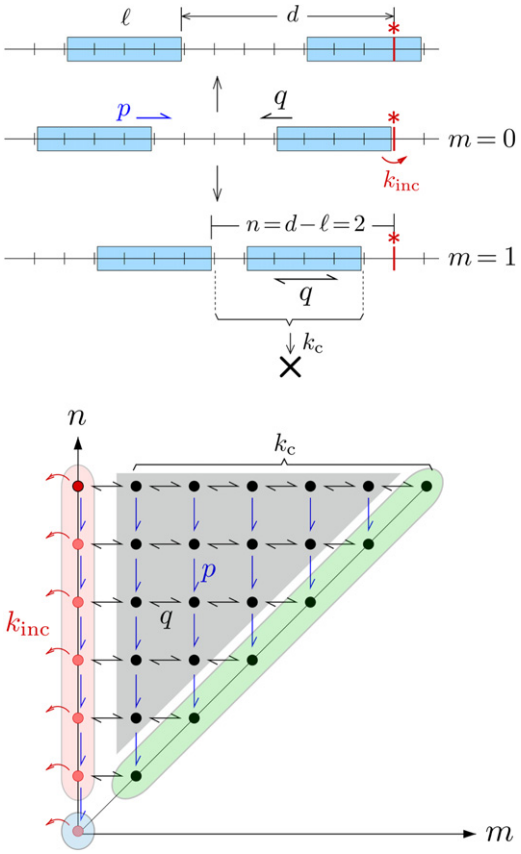

Figure 2. (Top panel) Local stochastic model for RNAP backtracking. (Bottom panel) State-space of RNAP immediately following a wrong nucleotide addition. When this happens, assume that the open distance to the second RNAP is $n(t=0) \equiv N$. While in a backtracking state, the leading RNAP diffusively hops forward and backward with rate q. The trailing RNAP hops forward with rate p. There are two mechanisms to escape from the backtracking state: incorporation of the wrong nucleotide with rate $k_{\rm inc}$ while in states $(m=0,n)$, and cleavage with rate $k_{\rm c}$ while in states $(1 \leqslant m \leqslant n)$. In the diagram, the states indicated by black dots can undergo cleavage, while those indicated by red dots can incorporate the error.

that drive the dynamics of the $n = N-1$ chain, and so on.

By defining the Laplace transform $\tilde{P}_n(m,s) = \int_0^\infty \mathrm{e}^{-st} P_n(m,t)\mathrm{d}t$ and taking the Laplace transform of equation (2), we first find $\tilde{P}_N(1,s)$ in terms of $\tilde{P}_N(0,s)$ and successively substitute into equation (1) to find for $0 \leqslant m \leqslant N$

$$\tilde{P}_N(m,s) = \frac{D_{m-1}}{q^m}\left[\tilde{P}_N(0,s) - \sum_{k=0}^{m-1}\frac{q^{2k}}{D_{k-1}D_k}\right], \quad (5)$$

where the coefficients D_m obey

$$D_{m+1} = (s + k_{\rm c} + 2q + p)D_m - q^2 D_{m-1}, \quad m \geqslant 0. \quad (6)$$

To determine $\tilde{P}_N(0,s)$ and close the system, we apply the boundary condition at the end of the chain (equation (3)) to find $\tilde{P}_N(N,s) = q\tilde{P}_N(N-1)/(s + k_{\rm c} + q)$. Upon using equation (5) for $\tilde{P}_N(N,s)$ and $\tilde{P}_N(N-1,s)$, we find

$$D_{N-1}\left[\tilde{P}_N(0,s) - \sum_{k=0}^{N-1}\frac{q^2 k}{D_{k-1}D_k}\right] = \frac{q^2 D_{N-2}}{s + k_{\rm c} + q}\left[\tilde{P}_N(0,s) - \sum_{k=0}^{N-2}\frac{q^{2k}}{D_{k-1}D_k}\right], \quad (7)$$

from which we find $\tilde{P}_N(0,s)$ explicitly

$$\tilde{P}_N(0,s) = \frac{1}{D_{N-2}}\left[\frac{q^{2(N-1)}(s + k_{\rm c} + q)}{(s + k_{\rm c} + q)D_{N-1} - q^2 D_{N-2}}\right] + \sum_{k=0}^{N-2}\frac{q^{2k}}{D_{k-1}D_k}. \quad (8)$$

The recursion in D_m starts with $D_{-1} \equiv 1$, $D_0 = s + k_{\rm inc} + p + q$. To find an explicit expression for D_m, we use the generating function $G(z) \equiv \sum_{m=0}^\infty D_m z^m$ to convert equation (6) to

$$\frac{1}{z^2}\left[G(z) - D_0 - D_1 z\right] = \frac{A}{z}\left[G(z) - D_0\right] - q^2 G(z), \quad (9)$$

which is solved by

$$G(z) = \frac{D_0 + zD_1 - zAD_0}{q^2 z^2 - Az + 1} = \frac{D_0(1 - zA) + zD_1}{q^2(z_+ - z_-)}\left(\frac{1}{z - z_+} - \frac{1}{z - z_-}\right), \quad (10)$$

where $z_\pm > 0$ and $z_+ > z_-$:

$$z_\pm = \frac{(s + \lambda_{\rm c})}{2q^2}\left[1 \pm \sqrt{1 - \frac{4q^2}{(s + \lambda_{\rm c})^2}}\right]. \quad (11)$$

By using $(1 - z/z_\pm)^{-1} = \sum_{k=0}^\infty (z/z_\pm)^k$, we find the power series of $G(z)$ about $z = 0$ (or use the inverse Z-transform) to find

$$G(z) = \frac{D_0 - (D_1 - AD_0)z}{(z_+ - z_-)}\left[\frac{1}{z_-}\sum_{m=0}^\infty\left(\frac{z}{z_-}\right)^m - \frac{1}{z_+}\sum_{m=0}^\infty\left(\frac{z}{z_+}\right)^m\right], \quad (12)$$

and hence an explicit expression for D_m:

$$D_{m\geqslant 2} = \frac{D_0}{q^2(z_+ - z_-)}\left(\frac{1}{z_-^{m+1}} - \frac{1}{z_+^{m+1}}\right) + \frac{(D_1 - AD_0)}{q^2(z_+ - z_-)}\left(\frac{1}{z_-^m} - \frac{1}{z_+^m}\right). \quad (13)$$

We can substitute $\tilde{P}_N(0,s)$ from equation (8) into equation (5) and use the above expression for D_m to find an explicit solution to $\tilde{P}_N(m,s)$. The above results assume a fixed trailing RNAP but will be used to construct the full solution in the presence of

a forward-moving trailing RNAP. Nonetheless, this one-row ($n = N$) approximation provides a lower bound on the probability that the wrong nucleotide is incorporated.

2.2. Closing trailing particle

Since elongation is irreversible, the system is feed-forward; that is, the probabilities in the $n = N$ layer feed into the $n = N - 1$ layer, and so on. The probability flux from the n chain into each state $m \leqslant n - 1$ of the $n - 1$ chain is $\tilde{J}_{n-1}(m, s) = p\tilde{P}_n(m, s)$. Thus, the probabilities within the $n - 1$ chain can be described by a recursion relation with an additional source of probability from the n layer:

$$q\tilde{P}_n(m+1,s) = \frac{D_m}{D_{m-1}}\tilde{P}_n(m,s) - \frac{q^m}{D_{m-1}}\sum_{k=0}^{m} D_{k-1}q^{-k}\tilde{J}_n(k,s), \quad (14)$$

where $\tilde{J}_n(k, s) = p\tilde{P}_{n+1}(k, s)$. Equation (14) can be easily recursed to find an explicit expression for $\tilde{P}_n(m, s)$ in the n layer:

$$\tilde{P}_n(m,s) = \frac{D_{m-1}}{q^m}\left[\tilde{P}_n(0,s) - \sum_{\ell=0}^{m-1} q^\ell \tilde{Q}_n(\ell,s)\right], \quad (15)$$

where

$$\tilde{Q}_n(\ell,s) = \frac{q^\ell}{D_\ell D_{\ell-1}}\sum_{k=0}^{\ell} D_{k-1}q^{-k}\tilde{J}_n(k,s). \quad (16)$$

We now Laplace-transform the boundary condition in equation (3) to find

$$\tilde{P}_n(n,s) = \frac{\tilde{J}_n(n,s) + q\tilde{P}_n(n-1,s)}{s+q+k_c}. \quad (17)$$

After using equation (15) for $\tilde{P}_n(n, s)$ and $\tilde{P}_n(n - 1, s)$ in equation (17), we can explicitly solve for

$$\tilde{P}_n(0,s) = \frac{q^n\tilde{J}_n(n,s) - q^2D_{n-2}\sum_{\ell=0}^{n-2} q^\ell\tilde{Q}_n(\ell,s)}{(s+q+k_c)D_{n-1} - q^2D_{n-2}} + \frac{(s+q+k_c)D_{n-1}\sum_{\ell=0}^{n-1} q^\ell\tilde{Q}_n(\ell,s)}{(s+q+k_c)D_{n-1} - q^2D_{n-2}}, \quad (18)$$

which we can use in equation (15) to find an explicit expression for $\tilde{P}_n(m, s)$. Note that $\tilde{P}_n(m, s)$ depends on $\tilde{Q}_n(\ell, s) \propto \tilde{J}_n = p\tilde{P}_{n+1}$, the probabilities in the layer immediately above it.

2.3. Outcome probabilities and times

With the Laplace-transformed probabilities derived, we can calculate the probabilities that the erroneous NTP is incorporated or cleaved. The probability that the RNAP incorporates the wrong nucleotide by time t can be calculated by time-integrating the probability flux

$$P_{\text{inc}}(t) = k_{\text{inc}}\sum_{n=0}^{N}\int_0^t P_n(m=0,t')\mathrm{d}t'. \quad (19)$$

The final probability of wrong nucleotide incorporation is $P_{\text{inc}}(\infty) = k_{\text{inc}}\sum_{n=0}^{N}\tilde{P}_n(m=0,s=0)$, while the total probability of cleaving is $P_c(\infty) = 1 - P_{\text{inc}}(\infty)$.

While our results apply for the times after an RNAP enters a backtracking state, we can extend them by weighting the overall incorporation probability P_{inc} by the probability $1 - P^*$ that leading RNAP enters the backtracking state after associating with a noncognate nucleotide. The unconditional probability P'_{inc} that a mismatched nucleotide is incorporated is thus

$$P'_{\text{inc}} = P_{\text{inc}}(1 - P^*) + P^*, \quad (20)$$

and the total error correction probability is $P'_c = 1 - P'_{\text{inc}}$.

Conditioned on incorporation of a mismatched nucleotide within the backtracking state, we can also define the density of incorporation times as $w(t) = k_{\text{inc}}P_n(m=0,t)/P_{\text{inc}}(\infty)$ and find the moments of the conditioned incorporation time [53]

$$\mathbb{E}[T^\sigma_{\text{inc}}] = \frac{(-1)^\sigma k_{\text{inc}}}{P_{\text{inc}}(\infty)}\left[\frac{\partial^\sigma}{\partial s^\sigma}\sum_{n=0}^{N}\tilde{P}_n(m=0,s)\right]_{s=0}. \quad (21)$$

Similarly, the moments of the times to cleavage (and correction of the misincorporated nucleotide), conditioned on cleavage are

$$\mathbb{E}[T^\sigma_c] = \frac{(-1)^\sigma k_c}{P_c(\infty)}\left[\frac{\partial^\sigma}{\partial s^\sigma}\sum_{n=0}^{N}\sum_{m=1}^{n}\tilde{P}_n(m,s)\right]_{s=0}. \quad (22)$$

Finally, the resolution time-the time for the system to either cleave or incorporate the mismatched nucleotide (conditioned on starting from the backtracking state)-obeys

$$\mathbb{E}[T^\sigma] = (-1)^\sigma k_{\text{inc}}\left[\frac{\partial^\sigma}{\partial s^\sigma}\sum_{n=0}^{N}\tilde{P}_n(m=0,s)\right]_{s=0} + (-1)^\sigma k_c\left[\frac{\partial^\sigma}{\partial s^\sigma}\sum_{n=0}^{N}\sum_{m=1}^{n}\tilde{P}_n(m,s)\right]_{s=0}. \quad (23)$$

3. Results and discussion

Henceforth, we will nondimensionalize time by $1/q$ and measure all rates in terms of q. In figures 3(a) and (b), we use equation (19) to plot the final incorporation probability $P_{\text{inc}}(\infty)$ as a function of the incorporation rate k_{inc} and the initial RNAP separation N for different values of the trailing RNAP elongation rate p.

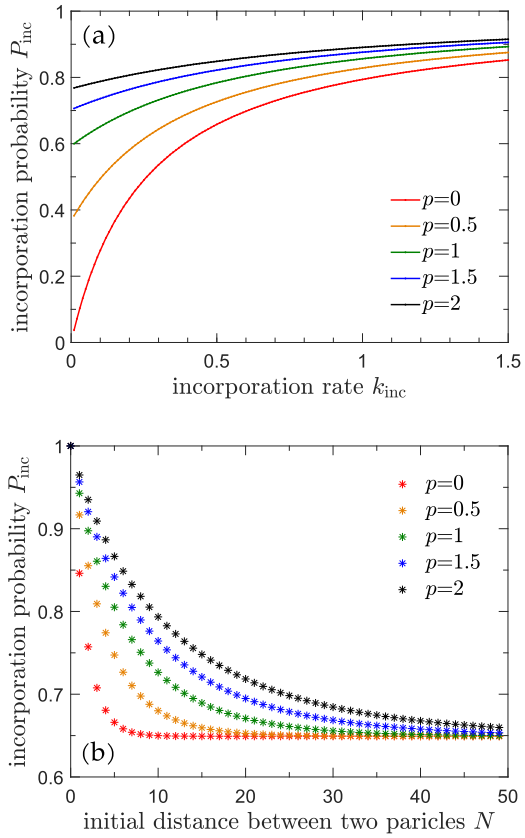

Figure 3. (a) Incorporation probability as a function of the error incorporation rate $k_{\rm inc}$. We measure all rates in terms of q, set $k_{\rm c} = 0.1$, and explore different values of p. (b) Incorporation probability $P_{\rm inc}(\infty)$ as a function of initial separation N. As N is decreased, the incorporation probability increases because the backtracking RNAP has less opportunity to cleave since it is more confined and spends more time abutted against the misincorporation site ($m = 0$). For both plots, the initial RNAP separation $N = 6$.

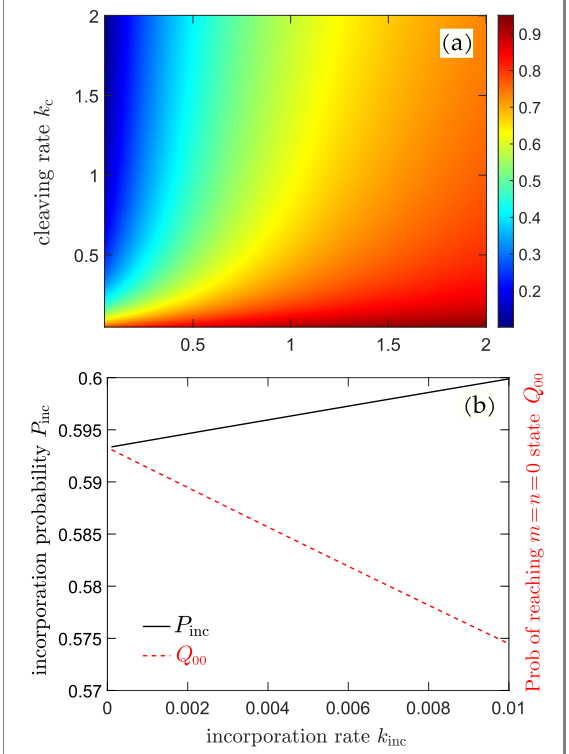

Figure 4. (a) Incorporation probability of the full model, plotted as a density with $p = 1$ and an initial gap size $N = 6$. $P_{\rm inc}$ is monotonically decreasing (increasing) with increasing $k_{\rm c}$ ($k_{\rm inc}$). The iso-probability lines are approximately quadratic in $k_{\rm inc}$. (b) For illustration, we keep $p = 1$, set $k_{\rm c} = 0.1$, and plot in the small $k_{\rm inc}$ limit, the probability of incorporation $P_{\rm inc}$ and the probability Q_{00} of reaching the compressed $m = n = 0$ state. Note that $P_{\rm inc}$ approaches a finite value as $k_{\rm inc} \to 0$ because the $m = 0$ states constitute a 'kinetic trap' if the trailing RNAP abuts against the leading RNAP.

Although $P_{\rm inc}(\infty) \propto k_{\rm inc}$, it increases sublinearly with $k_{\rm inc}$ (figure 3(a)) because random diffusion mitigates the incorporation by distributing the RNAP away from the $m = 0$ incorporation site. Nonetheless, as $k_{\rm inc}$ increases, the RNAP is more likely to incorporate the error. For a fixed $k_{\rm inc}$, having a faster elongation rate p yields higher incorporation probability since there is effectively less time for the leading RNAP to cleave the erroneous nucleotide.

Figure 3(b) shows that $P_{\rm inc}(\infty)$ converges to the common value $P_{\rm inc}(\infty) \approx 0.65$ as $N \to \infty$. This corresponds to an infinitely far trailing RNAP that will not influence error correction of the leading RNAP. Note that $P_{\rm inc}(\infty)$ reaches the asymptotic value 0.65 faster for smaller p. In all cases, the final error incorporation probability increases with RNAP translocation rate p and can be thought of as a trailing RNAP 'pushing' a backtracking-state (leading) RNAP to incorporate the error. In our model, the trailing RNAP serves only to restrict motion of the leading RNAP through steric exclusion and does not influence the forward elongation rate of the leading RNAP. Thus, the trailing RNAP pushes the leading RNAP not by a direct contact force with the leading RNAP, but by reducing the entropy of the backtracking RNAP ahead of it and increasing its chance of incorporating the erroneous nucleotide.

In figure 4(a), we fix $p = 1$, set the initial gap size $N = 6$, and plot $P_{\rm inc}(\infty)$ as a function of the cleavage rate $k_{\rm c}$ and the incorporation rate $k_{\rm inc}$. In figure 4(b) we show that the limiting behavior of $P_{\rm inc}(\infty) \nrightarrow 0$ as $k_{\rm inc} \to 0$. We define $Q_{00} \equiv k_{\rm inc}\tilde{P}_0(m = 0, s = 0) = p\tilde{P}_1(m = 0, s = 0)$ as the probability that the trailing RNAP contacts the leading RNAP at the realignment position (the probability that the $m = n = 0$ 'compressed' state is reached). As also shown in figure 4(b), $Q_{00} \nrightarrow 0$ as $k_{\rm inc} \to 0$. Since in the $m = n = 0$ state, the only way to escape from the backtracking state is through incorporation, $P_{\rm inc}(\infty) \geqslant Q_{00}$ because incorporation is not limited to occur from the $m = n = 0$ state. As $k_{\rm inc} \to 0$, we expect that incorporation can occur only when cleavage becomes impossible, which is the case in the $m = n = 0$ state. In the $m = n = 0$ state, the only way to escape backtracking is through incorporation. Therefore, as shown in figure 4(b), $P_{\rm inc} \to Q_{00}$ as

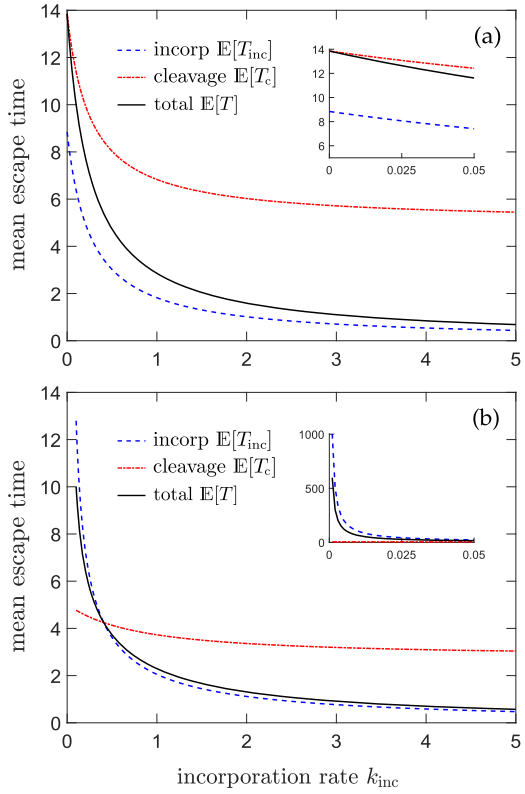

Figure 5. Mean times of incorporation ($\mathbb{E}[T_{\rm inc}]$ blue dashed), mRNA cleavage ($\mathbb{E}[T_{\rm c}]$ solid red), and unconditional ($\mathbb{E}[T]$ solid black). (a) Mean exit times in a system in which the trailing RNAP is stationary ($p = 0$) with fixed $n = N = 6$. (b) Mean exit times when the trailing RNAP advances (the full problem with $p = 1$) with a starting distance of $N = 6$. For both scenarios, we used $k_{\rm c} = 0.1$.

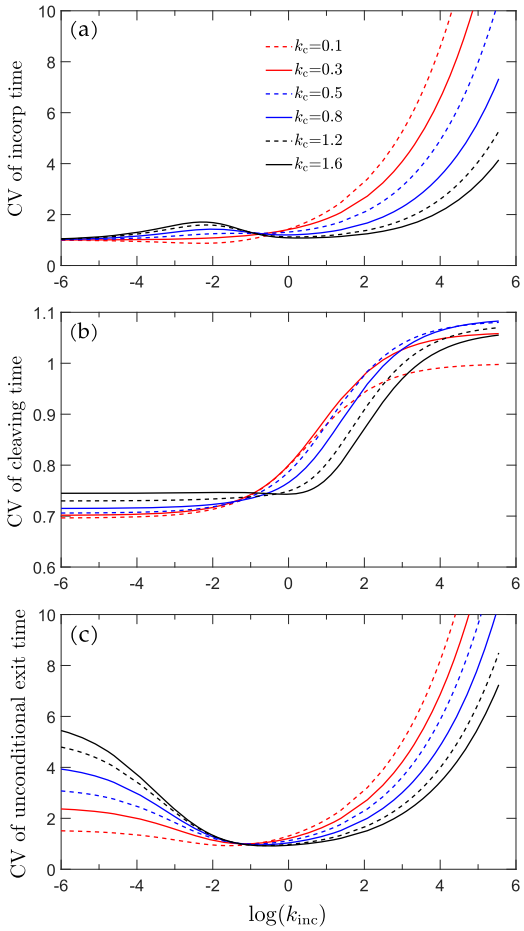

Figure 6. (a) The incorporation-time CV plotted as a function of $k_{\rm inc}$ for $p = 1$, $N = 6$, and various values of $k_{\rm c}$. (b) The CV of cleavage times. (c) General (unconditioned) escape times. These results show that the backtracking state (or 'pauses') can exhibit non-Poissonian waiting times.

$k_{\rm inc} \to 0$. This contact probability Q_{00} decreases when $k_{\rm c}$ increases or p decreases as the $m = n = 0$ state becomes less likely.

In figures 5(a) and (b) we use equations (21)–(23) to plot the mean backtracking-state escape times (first passage times), conditioned on incorporation, cleavage, or neither. When the trailing RNAP is stationary (figure 5(a)), the mean escape time conditioned on cleaving is always greater than the mean escape time conditioned on incorporation. Since the probability of incorporation vanishes as $k_{\rm inc} \to 0$, the unconditioned mean escape time approaches the mean time to cleave in this limit. In the inset of figure 5, we see that both the conditioned and unconditioned mean exit times remain finite as $k_{\rm inc} \to 0$ because the system can always escape by cleaving when the trailing RNAP is fixed.

We find qualitatively different behavior of mean escape times for the full model in which the trailing RNAP is allowed to advance. Figure 5(b) shows the conditioned and unconditioned mean exit times for a trailing RNAP with elongation rate $p = 1$. Here, the mean cleavage time is smaller than the mean incorporation time if $k_{\rm inc}$ is sufficiently small. For $k_{\rm inc} \to 0$, as shown in the inset, both the unconditional mean exit time and the mean incorporation time diverge. This divergence arises since occupation of the $m = n = 0$ state becomes more likely and the mean incorporation time from this state scales as $1/k_{\rm inc}$.

As the incorporation rate $k_{\rm inc}$ increases, the unconditioned mean exit time approaches the mean incorporation time, which decreases since it becomes increasingly likely for the leading particle to incorporate the erroneous nucleotide.

In principle, all moments of exit times can be directly computed from the s-dependence of $\tilde{P}_n(m, s)$ and equations (21)–(23). Here, we will simplify matters and consider only the coefficient of variation (CV) of the exit times

$$\mathrm{CV} = \frac{\sqrt{\mathbb{E}\left[(T - \mathbb{E}[T])^2\right]}}{\mathbb{E}[T]}. \tag{24}$$

These CVs involve only the first and second moments of the escape times and represent simple metrics that measure their deviation from those of Poisson processes for which CV = 1. Where appropriate, we substitute $\mathbb{E}[T_{\rm inc}]$ or $\mathbb{E}[T_{\rm c}]$ for $\mathbb{E}[T]$ above.

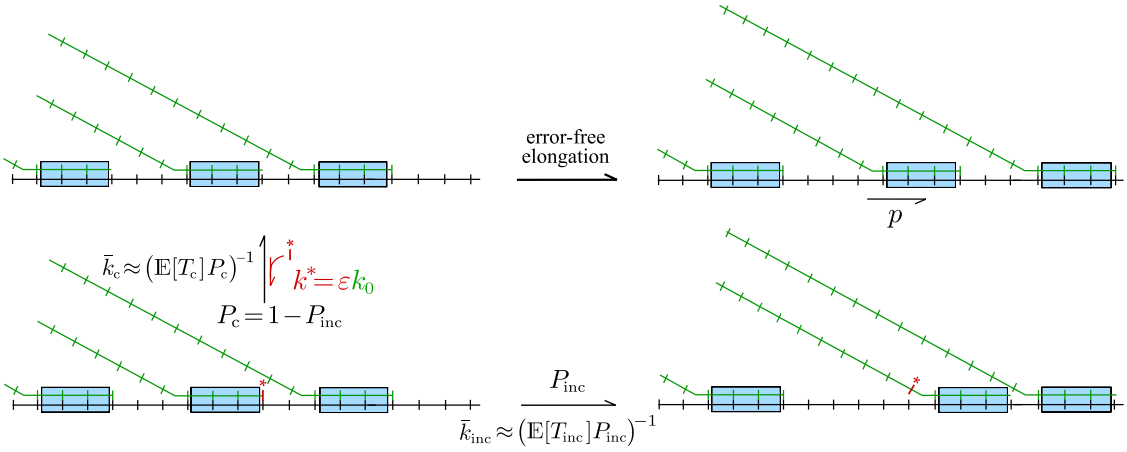

Figure 7. Effective lattice model incorporating two-particle results. The loading of a correct nucleotide occurs at rate k_0, while the loading of an incorrect nucleotide occurs at rate $k^* = \varepsilon k_0$. Once this occurs, the system enters the backtracking state. We assume that cleavage occurs before the leading RNAP diffuses too far back so that we can ignore the local diffusive motion during this time. Ultimately, the leading RNAP will error-correct by cleaving the nascent erroneous 3′ end, reverting back to the pre-backtracking state (upper left) with probability P_c and effective rate $\bar{k}_c \approx 1/(\mathbb{E}[T_c]P_c)$. Alternatively, with probability P_{inc}, the erroneous nucleotide gets incorporated with effective rate $\bar{k}_{inc} \approx 1/(\mathbb{E}[T_{inc}]P_{inc})$. To simulate this model, note that P_{inc}, P_c, $\mathbb{E}[T_{inc}]$, and $\mathbb{E}[T_c]$ depend implicitly on the distance N between the trailing and leading RNAPs at the instant the backtracking state is entered.

The escape-time CVs are plotted as functions of $\log(k_{inc})$ in figures 6(a)–(c) for $N = 6, p = 1$, and various k_c. The CV of the incorporation times shown in figure 6(a) indicates a Poisson process in the $k_{inc} \to 0$ limit as incorporation becomes a rare event. After peaking at an intermediate k_{inc}, the incorporation-time CV diverges as $\sqrt{k_{inc}}$ in the $k_{inc} \to \infty$ limit. Figure 6(b) shows a cleaving-time CV that is less than one for small k_{inc}, illustrating that cleaving can occur from multiple, connected states. For large k_{inc} and fixed k_c, the CV remains near one (see (iii) below). The asymptotic limits in (iv) below are not depicted in (b). Finally, the unconditioned exit time CV shown in figure 6(c) indicates a large CV for small k_{inc} that approaches the Poisson limit before increasing again at large k_{inc}. These results for the different waiting times indicate non-Poissonian behavior in RNAP pausing as was found by Voliotis *et al* under a different stochastic model [34].

The limiting behaviors of these CVs can be more simply understood and approximated by considering a toy model consisting of only two states: (1) an effective boundary $m = 0$ state that can immediately incorporate the error (with rate k_{inc}) and (2) an effective interior $m > 0$ state that allows cleavage at rate k_c. By lumping each of these two classes of states into a single state and labeling their probabilities by $P_0(t)$ and $P_1(t)$, respectively, we can explicitly find $\tilde{P}_0(s) = (s + k_c + q)/[(s + k_{inc} + q)(s + k_c + q) - q^2]$ and $\tilde{P}_1(s) = q/[(s + k_{inc} + q)(s + k_c + q) - q^2]$, where the diffusive hopping rate q in this simplified model becomes the inter-state transition rate. The initial condition for this toy model is $P_0(t = 0) = 1$, from which we find.

(a). The incorporation-time CV $\to 1$ as $k_c \to \infty$. When $k_c \to \infty$, incorporation is a rare process rate-limited by k_{inc}, leading to an effectively single Markovian (Poisson) step;

(b). The incorporation-time CV diverges as $\sqrt{k_{inc}}$ for $k_{inc} \to \infty$ (as shown in figure 6(a)). When k_{inc} is large, incorporation is less rare and its statistics are strongly affected by backtracking steps;

(c). The cleavage-time CV $\to 1$ when either k_c or $k_{inc} \to \infty$ and the other is large compared to q. By construction, the $m = 0$ state in our model precludes cleaving. Thus, in the limit $k_{inc} \to \infty$, cleavage occurs only after a single-step Poissonian transition to the $m = 1$, rate-limited by the backtracking hopping rate q;

(d). The CV of the cleavage times $\sim \sqrt{\frac{1+(k_c/k_{inc})^2}{(1+k_c/k_{inc})^2}}$ when $k_c, k_{inc} \to \infty$ with k_c/k_{inc} fixed. For example, if $k_c = k_{inc} \to \infty$, the cleavage-time CV $\sim 1/\sqrt{2}$;

(e). The CV of the overall (unconditioned) exit time diverges as $\sqrt{2k_{inc}}/k_c$ when $k_{inc} \to \infty$;

(f). The CV of the overall exit time ~ 1 as k_c, $k_{inc} \to \infty$ with k_c/k_{inc} fixed. In this limit, the dominant contribution to the exit time is cleavage that is limited by q, as in case (c) above.

The predictions of the CV from this toy model conform to limiting results of the full model shown in figure 6. Thus, these limiting behaviors are independent of finite RNAP spacing N. The CVs provide insight into the statistics of the exit times of a backtracking state and will be useful in developing effective multi-RNAP exclusion models, such as the modified TASEP described in figure 7, that can naturally allow for successive and/or multiple backtracking RNAPs. In particular, for regimes under which the CVs or first passage times are not too different from unity, we

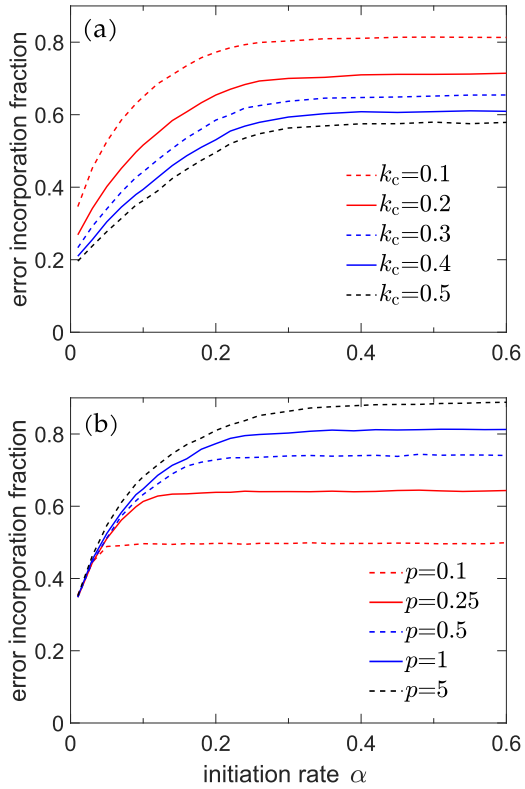

Figure 8. Preliminary proof-of-principle simulations of the effective lattice model defined in figure 7. We set the RNAP size $\ell = 1$, transcript length $L = 100$, termination rate $\beta = 0.5$, backtracking hopping rate $q = 1$, nucleotide association rate $k_0 = 1$, noncognate nucleotide association rate $k^* = \varepsilon k_0 = 0.02$, and plot the fraction of pauses that lead to nucleotide misincorporation as a function of initiation rate α. (a) Incorporation fraction as a function of α with fixed $p = 1$ and various values of k_c. (b) Fraction of incorporated errors as a function of α with fixed $k_c = 0.1$ and different values of the elongation rate p.

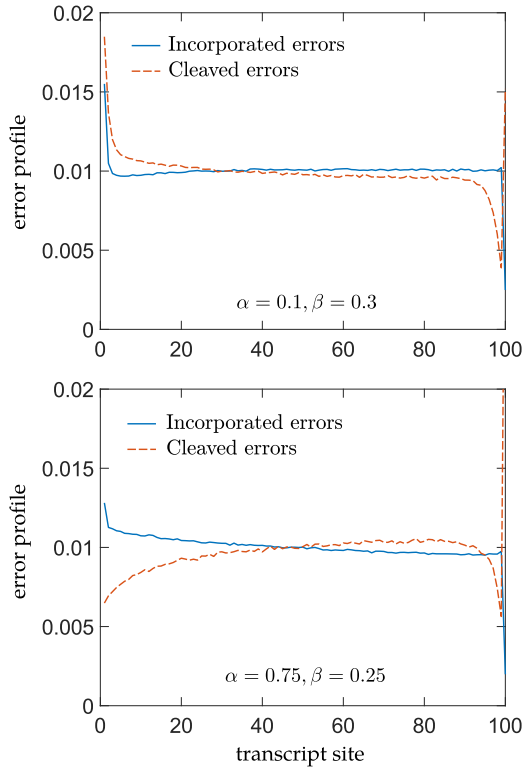

Figure 9. Simulated error profiles for $\ell = 1$, transcript length $L = 100$, elongation rate $p = 1$, backtracking hopping rate $q = 1$, nucleotide association rate $k_0 = 1$, noncognate nucleotide association rate $k^* = \varepsilon k_0 = 0.02$, and $k_c = k_{inc} = 0.1$. For both (a) $\alpha = 0.1, \beta = 0.3$ and (b) $\alpha = 0.75, \beta = 0.25$, the error profiles are quite flat.

may construct a coarse-grained model in which overall cleavage and incorporation are defined by constant effective rates as shown in figure 7.

While quantitative experimental data on the spatial profile of transriptional errors do not seem readily available. New sequencing protocols have been developed to interrogate the error 'spectra', that is, how frequently a specific DNA codon is is transcribed into a non-corresponding RNA codon. A particular finding is a G → A substitution bias in certain bacteria [54]. Nonsense errors, which are a small fraction of the total transcription error, were also found to increase in frequency closer to the normal stop codon (3′ end). However, the overall error rate seems to be fairly constant across genomes, with no apparent large scale spatial pattern [14, 55]. Moreover, the error rate seems to be insensitive to the level of gene expression.

We performed simulations of our toy effective model and show under what conditions does it predict an incorporated error rate that is rather insensitive to expression level. Figure 8 shows results from preliminary Monte-Carlo simulations of the effective model depicted in figure 7. For simplicity, we set $\ell = 1$, a transcript length $L = 100$, a termination rate $\beta = 0.5$, $k_{inc} = 0.1$, $k_0 = 1$ (arbitrary units), and $\varepsilon = 0.02$. We plot the fraction of pauses that lead to incorporated nucleotide errors. In (a), we set $p = 1$, and vary k_c, while in (b), we set $k_c = 0.1$ and vary p. In both cases, we see that for $\beta = 0.5$, increasing the initiation rate α beyond 0.2 does not lead to significantly higher RNAP densities, and thus higher incorporation probabilities. The competition between cleaving and incorporation is evident in (a) while the pushing effect (larger p leading to higher probability of incorporation) is shown in (b). In the large α limit of these simulations, about two pauses occurred in the production of each transcript. We have tested a smaller termination rate $\beta = 0.3$ and still find the expected number of errors per transcript to be fairly constant for $\alpha \gtrsim 0.2$.

In figures 9(a) and (b) we plot the mean profile of the number of repairs (cleaved errors) and the incorporated errors. We see that the density incorporated errors (what is experimentally observed) is fairly uniform except for at the initiation and termination sites. This property holds for both $\alpha = 0.1, \beta = 0.3$ (which would normally be a low-density phase in the standard TASEP) as well as $\alpha = 0.75, \beta = 0.25$ (which would normally be in the high-density phase in the standard TASEP).

To motivate the predictions observed from our toy model, note that in our model backtracking can only be induced with a mismatched nucleotide is loaded. We have implemented this mechanism by assigning a small probability that an RNAP enters the backtracking state with each elongation step, as described by equation (20). Thus, each transcript will suffer the same number of backtracking events, regardless of how much time the associated RNAP take to produce the transcript. In such a model, the only effect of RNAP density (N) would be on the probability of error incorporation. Under the parameters explored, we find (i) a fairly uniform error profile, and (ii) a total transcript error rate that is insensitive to initiation rate $\alpha \gtrsim 0.2$ when $\beta = 0.5$.

4. Summary and conclusions

RNAP backtracking is an important mechanism for transcription fidelity [20, 21, 42] as it is an intermediate step before cleavage of a misincorporated nucleotide. To study this process, we derived a stochastic model describing the interactions between two processing RNAP enzymes after the leading one has incorporated an erroneous nucleotide and transitioned into a backtracking state. We used Laplace transforms to formally solve the three-parameter stochastic model and found the probabilities for removing or incorporating the erroneous nucleotide. Our implicitly time-dependent analyses allow for easy computation of probabilities as well as conditional moments of times to error removal or incorporation.

Previous studies have concluded that the trailing RNAP will likely 'push' the leading backtracking RNAP forward, making it exit the backtracking state faster [20, 21, 42, 43, 51]. Such pushing mechanisms have also been observed in couple transcription-translation, in which a *ribosome* pushes the leading RNAP out of the backtracking state [56]. It is important to note that some pushing models also include the effects of *concerted* or cooperative motion of adjacent RNAPs that arise from a strong 'power stroke' [43]. In our analysis, we include only an entropic Brownian ratchet mechanism of pushing, but we allow for a dynamically closing gap starting from an initial gap distance N at the instant the leading RNAP entered the backtracking state. Our pushing mechanism reduces, over time, the states that allow cleavage, thereby biasing the system to more likely incorporate the noncognate nucleotide. However, the conditional times to cleavage, incorporation, or exiting the backtracking state depend more subtly on an advancing trailing RNAP. While the entropic pushing leads to a higher probability of incorporation, it also leads to a higher conditional mean time to incorporation, particularly at low values of k_{inc}. Pushing also leads to a significantly shorter conditional mean time of cleavage and a slightly higher unconditional mean time to leave the backtracking state, as shown in figures 5(a) and (b). This behavior indicates that without pushing, incorporation competes with cleavage, arising only when incorporation is fast. With reduced cleavage, conditional incorporation is more strongly governed by k_{inc}, which, if small, results in a large conditional incorporation time.

Finally, we simulated an effective toy TASEP-like model informed by our analyses model and computed quantities such as the expected incorporated errors per transcript and the distribution of errors along each transcript. These provide only a rudimentary mechanistic understanding of the effects of exclusion and error correction from backtracking. Measurements and inference of position-dependent error probabilities have been performed, but the errors are known to depend on other complicating factors such as the presence of cofactors and to be sequence-dependent (the 'error spectrum') [14]. To explore density-dependent error rates, one would have to first factor out nonsense and insertion/deletion errors, and errors that arise from sequence dependence. These more complex modifications can be incorporated in future exclusion-type lattice models and simulated.

Acknowledgments

This work was supported by Grants from the NIH through Grant R01HL146552 (TC), the Army Research Office through grant W911NF-18-1-0345 (TC), and the NSF through Grant DMS-1814364 (TC). The authors thank S K Lyons for feedback on the manuscript.

Data availability statement

No new data were created or analysed in this study.

Appendix A. Alternate calculation of mean times

Our method of solution requires solution of the recursion relations for the probabilities as a function of all the rate parameters and the Laplace-transformed time variable s. Thus, we explicitly carry all time dependence throughout the calculation in terms of s. In the end, we either set $s = 0$ to find probabilities or take derivatives with respect to s and then set $s \to 0$ to find moments of the escape times.

However, if we are only interested in the mean condition times to cleavage or incorporation, we can develop a simple coupled set of recursion relations that can be easily evaluated numerically.

A.1. Conditional mean times for a stationary trailing RNAP

First, consider the case where the trailing RNAP is stationary—we treat the more general case of an advancing trailing particle in the next subsection. For an initial gap N between the trailing RNAP and the backtracking RNAP, we want to find the expected time $\mathbb{E}[T_c]$ for the backtracking particle to cleave (correct) the error, and the expected time $\mathbb{E}[T_c]$ to incorporate the error given that cleavage or incorporation, respectively, occurs.

A static trailing particle means that the system stays in the first row of the state diagram in figure 2 (bottom panel). Let us label the states of the top row from left to right as $B_0, B_1, \ldots, B_N$. B_0 corresponds to the initial state where the leading RNAP has just added a wrong nucleotide but has not yet incorporated it. B_N corresponds to the state where the leading RNAP has backtracked a distance N and abuts the trailing RNAP.

If the RNAP incorporates the error while in state B_0, then we denote this state as B_{-1}; if the RNAP cleaves the error from state B_i, $0 < i \leqslant N$, then we denote this state as B_{N+1}. Note that B_{-1} and B_{N+1} represent absorbing states associated with error incorporation and error correction respectively.

Define $v_k = \mathbf{P}_k(X_T = B_{-1})$ as the probability that the system reaches state B_{-1} given that it started in state B_k. These probabilities satisfy the recursion relations

$$\begin{aligned} v_1 &= (1+\gamma)v_0 - \gamma \\ v_i &= \frac{v_{i-1}}{\beta} - v_{i-2}, \quad 2 \leqslant i \leqslant N \\ v_N &= \frac{q}{q+k_c} v_{N-1}, \end{aligned} \tag{25}$$

where $\gamma = k_{inc}/q$, $\beta = q/(2q+k_c)$, $v_{-1} = 1$, and $v_{N+1} = 0$. One can show that the solution to equations (25) is given by

$$\begin{aligned} v_i &= C_i v_0 - F_i \\ C_i &= x_1 \zeta_+^i + x_2 \zeta_-^i \\ F_i &= x_3 \zeta_+^i + x_4 \zeta_-^i, \end{aligned} \tag{26}$$

where $\zeta_\pm = \frac{1 \pm \sqrt{1-4\beta^2}}{2\beta}$, $x_1 = \frac{\gamma+1-\zeta_-}{\zeta_+ - \zeta_-}$, $x_2 = 1 - x_1$, $x_3 = \frac{\gamma}{\zeta_+ - \zeta_-}$, and $x_4 = -x_3$. v_0 can be solved by plugging in the above expressions for v_i into the last equation in (25) which gives

$$v_0 = \frac{(q+k_c)F_N - qF_{N-1}}{(q+k_c)C_N - qC_{N-1}}. \tag{27}$$

One can check that equation (27) and $k_{inc}\tilde{P}(0, N, s = 0)$ (equation (21)) yield the same result when we set the elongation rate $p = 0$.

Next, we can study the mean escape time conditioned on incorporation. Recall that the conditional expectation of a random variable X given an event H, where $P(H) > 0$, is given by

$$\mathbb{E}[X|H] = \frac{\mathbb{E}[X \cdot 1_H]}{P(H)}. \tag{28}$$

We see directly from equation (28) that it is necessary to require $P(H) > 0$ for our equation to be well-defined (interested readers can read about the Borel–Kolmogorov paradox for the case $P(H) = 0$). Following this idea, we define $u_i = \mathbb{E}_i[T \cdot 1_A]$, where T is the time to reach one of the absorption states (i.e. unconditioned escape time), $\mathbf{1}_A$ is the indicator function for the event $A = \{X_T = B_{-1}\}$. The quantity we would like to find is $\mathbb{E}_0[T|A] = u_0/v_0$, which is the mean escape time conditioned on incorporation when the leading RNAP adds a wrong nucleotide. Since we have already solved for v_0, it suffices to find u_0. At each state B_j, there are rates for additional transitions depending on j. We can view these as competing Poisson processes. Suppose the rates are given by $r_{j_1}, \ldots, r_{j_d}$ at state B_j. Then the mean waiting time for the next move is $(\sum_{i=1}^d r_{j_i})^{-1}$. And the probability of choosing the move with rates r_{j_l} is simply $r_{j_l}/(\sum_{i=1}^d r_{j_i})$. Therefore, we obtain

$$\begin{aligned} \mathbb{E}_0[T \cdot 1_A] &= \frac{k_{inc}}{q+k_{inc}} \mathbb{E}_{-1}\left[\left(T + \frac{1}{q+k_{inc}}\right) \cdot 1_A\right] \\ &\quad + \frac{q}{q+k_{inc}} \mathbb{E}_1\left[\left(T + \frac{1}{q+k_{inc}}\right) \cdot 1_A\right] \\ \mathbb{E}_i[T \cdot 1_A] &= \frac{q}{2q+k_c} \mathbb{E}_{i-1}\left[\left(T + \frac{1}{2q+k_c}\right) \cdot 1_A\right] \\ &\quad + \frac{q}{2q+k_c} \mathbb{E}_{i+1}\left[\left(T + \frac{1}{2q+k_c}\right) \cdot 1_A\right] \\ \mathbb{E}_N[T \cdot 1_A] &= \frac{q}{q+k_c} \mathbb{E}_{N-1}\left[\left(T + \frac{1}{q+k_c}\right) \cdot 1_A\right]. \end{aligned} \tag{29}$$

After some algebra, we find

$$\begin{aligned} u_1 &= (1+\gamma)u_0 - \frac{v_0}{q} \\ u_{i+1} &= \frac{u_i}{\beta} - \frac{v_i}{q} - u_{i-1} \\ u_N &= \frac{q}{q+k_c} u_{N-1} + \frac{v_N}{q+k_c}. \end{aligned} \tag{30}$$

The above recursion relation can be solved analytically as

$$u_i = H_i u_0 - (G_i v_0 - K_i), \tag{31}$$

where

$$\begin{aligned} H_n &= C_n \\ G_n &= s_1 \zeta_+^n + s_2 \zeta_-^n + n s_3 \zeta_+^n + n s_4 \zeta_-^n \\ K_n &= j_1 \zeta_+^n + j_2 \zeta_-^n + n j_3 \zeta_+^n + n j_4 \zeta_-^n, \end{aligned} \tag{32}$$

and the coefficients s_i and j_i are given by

$$
\begin{aligned}
s_1 &= -\frac{\beta(-1+\beta(1+2\beta+\gamma))}{q(1-4\beta^2)^{3/2}}\\
s_2 &= -s_1\\
s_3 &= \frac{\beta(-1+\sqrt{1-4\beta^2}+2\beta(1+\gamma))}{q(2-8\beta^2)}\\
s_4 &= \frac{\beta(1+\sqrt{1-4\beta^2}-2\beta(1+\gamma))}{q(-2+8\beta^2)}\\
j_1 &= -\frac{\beta^2\gamma}{q(1-4\beta^2)^{3/2}}\\
j_2 &= -j_1\\
j_3 &= j_4 = \frac{\beta^2\gamma}{q-4q\beta^2}.
\end{aligned}
\tag{33}
$$

To find u_0, one can substitute equation (31) into the last equation of (30) to obtain

$$
\begin{aligned}
u_0 = &\frac{[C_N + q(G_N - G_{N-1}) + k_c G_N]v_0}{q(C_N - C_{N-1}) + k_c C_N}\\
&- \frac{[F_N + q(K_N - K_{N-1}) + k_c K_N]}{q(C_N - C_{N-1}) + k_c C_N}.
\end{aligned}
\tag{34}
$$

Similarly, we can find the mean escape time conditioned on cleaving. One can define $\bar{v}_i = 1 - v_i$ as the probability of starting from state B_i and eventually ending in state B_{N+1} (since there are only two absorbing states, a particle has to arrive at one of them). One can check that

$$
\begin{aligned}
\bar{v}_0 &= q\bar{v}_1 + (1 - q - k_{\text{inc}})\bar{v}_0\\
\bar{v}_i &= \frac{k_c\beta}{q} + \beta(\bar{v}_{i+1} + \bar{v}_{i-1})\\
\bar{v}_N &= \frac{k_c}{q+k_c} + \frac{q\bar{v}_{N-1}}{q+k_c}.
\end{aligned}
\tag{35}
$$

If we define $\bar{u}_i = \mathbb{E}_i[T \cdot 1_{A^c}]$, as the expected stopping time for the event $A^c = \{X_T = B_{N+1}\}$, we can show that $\bar{u}_i$ satisfies the same equations as u_i does with v_i changed to $\bar{v}_i$. The new recursion relation for $\bar{u}_i$ can be expressed as $\bar{u}_i = \bar{H}_i\bar{u}_0 - (\bar{G}_i v_0 - \bar{K}_i)$. As before, we have $\bar{H}_i = C_i$, $\bar{G}_i = -G_i$. $\bar{K}_i$ takes on a different form because the recursive relation for $\bar{K}_i$ yields one more root,

$$
\bar{K}_n = \bar{j}_1\zeta_+^n + \bar{j}_2\zeta_-^n + n\bar{j}_3\zeta_+^n + n\bar{j}_4\zeta_-^n + \bar{j}_5,
$$

where

$$
\begin{aligned}
\bar{j}_1 &= -\frac{\beta\left[1+\sqrt{1-4\beta^2}-2\beta(2\beta-\sqrt{1-4\beta^2}+\gamma)\right]}{2q(1-4\beta^2)^{3/2}}\\
\bar{j}_2 &= -\frac{\beta\left[-1+\sqrt{1-4\beta^2}+2\beta(2\beta+\sqrt{1-4\beta^2}+\gamma)\right]}{2q(1-4\beta^2)^{3/2}}\\
\bar{j}_3 &= \bar{j}_4 = -\frac{\beta^2\gamma}{q-4q\beta^2}\\
\bar{j}_5 &= \frac{\beta}{q-2q\beta}.
\end{aligned}
$$

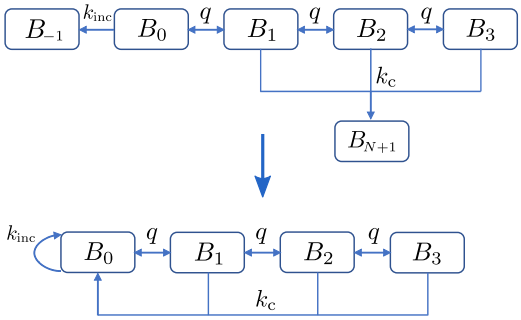

Figure 10. Rewiring of states for computing mean exit times. The (top panel) shows the original transition of states. Recall that B_{-1} represents incorporation of error, and B_{N+1} represents cleaving the error. In the (bottom panel), all transitions to absorbing states (B_{-1} and B_{N+1}) are rewired to the initial state B_0.

We then find $\bar{u}_0$

$$
\begin{aligned}
\bar{u}_0 = &\frac{1}{(q+k_c)C_N - qC_{N-1}}\Big[1 - C_N v_0 + F_N\\
&- q(\bar{G}_{N-1}v_0 - \bar{K}_{N-1}) + (q+k_c)(\bar{G}_N v_0 - \bar{K}_N)\Big].
\end{aligned}
$$

With u_0 and $\bar{u}_0$ given, we are able to calculate the unconditioned mean escape time T_u which is defined by

$$
T_u = v_0\mathbb{E}_i[T|1_A] + \bar{v}_0\mathbb{E}_i[T|1_{A^c}] \equiv u_0 + \bar{u}_0. \tag{37}
$$

One can also apply the same arguments to T_u as we used for u_i and $\bar{u}_i$ and seek $\mathbb{E}_i[T]$ instead of $\mathbb{E}_i[T \cdot 1_A]$.

Another way to calculate T_u was introduced by Hill [57]. He showed that the unconditioned mean escape time can be calculated if we consider the steady state in a transformed network without absorbing states. The transformed network is obtained by rewiring the transitions to absorbing states to the initial state in the original network. For instance, the maximum backtracking depth is set to be $N = 3$ in figure 10, and all transitions to absorbing states are rewired to the initial state.

The probability distribution of the stationary state of the rewired network can be found as $P_i = P_N L_{N-i}$, where $L_i = x_1'\zeta_+^i + x_2'\zeta_-^i$, $x_1' = (\lambda + 1 - \zeta_-)/(\zeta_+ - \zeta_-)$, $x_2' = 1 - x_1'$, $\lambda = k_c/q$, and

$$
P_N = \frac{1}{x_1'\frac{1-\zeta_+^{N+1}}{1-\zeta_+} + x_2'\frac{1-\zeta_-^{N+1}}{1-\zeta_-}}.
$$

The unconditioned mean escape time is given by

$$
T_u = \frac{1}{P_0 k_{\text{inc}} + (1-P_0)k_c}. \tag{38}
$$

One should note that we can get T_u for free by using equation (37) if we have the conditional incorporation time $\mathbb{E}_i[T|1_A]$, the conditional cleavage time $\mathbb{E}_i[T|1_{A^c}]$, the incorporation and cleavage probability v_0, and $\bar{v}_0$. However, one cannot recover the

conditional mean times $\mathbb{E}_i[T|1_A]$ and $\mathbb{E}_i[T|1_{A^c}]$ even if we know T_u, v_0, and $\bar{v}_0$ because essentially, we are trying to solve $xp + y(1-p) = c$ for both x and y [57], which does not have a unique solution.

A.2. Mean conditional times for a trailing RNAP that advances

To derive the incorporation probability when the trailing RNAP is moving forward with elongation rate p, we use figure 2 to build our solution. Let $v(m,n)$ be the probability of incorporating the error, given that the RNAP starts at state (m,n). Note that by definition, $v(i,j)$ only makes sense for $0 \leqslant i \leqslant j \leqslant N$, where N is the maximum backtracking depth, which is also the distance between the trailing and leading RNAP when the backtracking dynamics first started.

As a boundary condition, we have $v(0,0) = 1$. This is because when the leading RNAP is at the realignment position and there is no room for backtracking, it can only incorporate the error and move forward. Suppose we now have $v(i,j)$ for all $0 \leqslant i \leqslant j$; we can recursively build the solution for $v(i,j+1)$ for $0 \leqslant i \leqslant j+1$ via

$$v(0,j+1) = \frac{k_{\text{inc}} + pv(0,j) + qv(1,j+1)}{k_{\text{inc}} + p + q},$$

$$v(i,j+1) = \frac{qv(i-1,j+1) + qv(i+1,j+1) + pv(i,j)}{k_c + 2q + p} \quad \text{for all } 1 \leqslant i \leqslant j$$

$$v(j+1,j+1) = \frac{qv(j,j+1)}{k_c + q}. \tag{39}$$

We can use similar method as the previous section and study (numerically) the mean escape times. Let $u(i,j) = \mathbb{E}_{ij}[T \cdot 1_A]$, where T is the time to reach one of the absorption states, $\mathbf{1}_A$ is the indicator function for the event $A = \{X_T = B_{-1}\}$ and the subscript ij represents the initial state (m,n), with $0 \leqslant j \leqslant N$ and $0 \leqslant i \leqslant j$. The stochastic equations are given by

$$\begin{aligned}
u(0,0) &= \frac{1}{k_{\text{inc}}},\\
u(0,j) &= \frac{k_{\text{inc}}}{(k_{\text{inc}} + q + p)^2} + \frac{pu(0,j-1) + qu(1,j)}{k_{\text{inc}} + q + p}\\
&\quad + \frac{pv(0,j-1)}{(k_{\text{inc}} + q + p)^2} + \frac{qv(1,j)}{(k_{\text{inc}} + q + p)^2},\\
u(i,j) &= \frac{qu(i+1,j)}{2q + p + k_c} + \frac{qv(i+1,j)}{(2q + k_c + p)^2}\\
&\quad + \frac{qu(i-1,j)}{2q + p + k_c} + \frac{qv(i-1,j)}{(2q + k_c + p)^2}\\
&\quad + \frac{pu(i,j-1)}{2q + p + k_c} + \frac{pv(i,j-1)}{(2q + k_c + p)^2},\\
u(j,j) &= \frac{qu(j-1,j)}{q + k_c} + \frac{qv(j-1,j)}{(q + k_c)^2}.
\end{aligned} \tag{40}$$

The derivation for $\tilde{u}(i,j) = \mathbb{E}_{ij}[T \cdot 1_{A^c}]$ is similar. The linear system given in equations (39) and (40) can be easily solved since the size of the matrix in the linear system is on the order of the typical gap size between RNAPs during transcription. The mean time for a backtracking polymerase to incorporate the wrong nucleotide is $u(0,N)/v(0,N)$ when the initial distance from the trailing polymerase is N. The corresponding mean time for a backtracking polymerase to cleave the wrong nucleotide is $\tilde{u}(0,N)/\tilde{v}(0,N)$ and the unconditioned mean escape time is $u(0,N) + \tilde{u}(0,N)$. The above analyses provides an alternative methods for computing mean exit times and have been verified against the direct method presented in the main text.

ORCID iDs

Tom Chou https://orcid.org/0000-0003-0785-6349

References

[1] Derrida B, Janowsky S A, Lebowitz J L and Speer E R 1993 Exact solution of the totally asymmetric simple exclusion process: shock profiles *J. Stat. Phys.* **73** 813–42

[2] MacDonald C T, Gibbs J H and Pipkin A C 1968 Kinetics of biopolymerization on nucleic acid templates *Biopolymers* **6** 1–25

[3] Derrida B, Evans M R, Hakim V and Pasquier V 1993 Exact solution of a 1D asymmetric exclusion model using a matrix formulation *J. Phys. A: Math. Gen.* **26** 1493

[4] Lakatos G and Chou T 2003 Totally asymmetric exclusion processes with particles of arbitrary size *J. Phys. A: Math. Gen.* **36** 2027–41

[5] Zia R K P, Dong J J and Schmittmann B 2011 Modeling translation in protein synthesis with TASEP: a tutorial and recent developments *J. Stat. Phys.* **144** 405

[6] Erdmann-Pham D D, Dao Duc K and Song Y S 2020 The key parameters that govern translation efficiency *Cell Syst.* **10** 183–92

[7] Lynch M 2011 The lower bound to the evolution of mutation rates *Genome Biol. Evol.* **3** 1107–18

[8] Lang G I and Murray A W 2008 Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae Genetics* **178** 67–82

[9] Zhu Y O, Siegal M L, Hall D W and Petrov D A 2014 Precise estimates of mutation rate and spectrum in yeast *Proc. Natl Acad. Sci. USA* **111** E2310–8

[10] Shaw R J, Bonawitz N D and Reines D 2002 Use of an *in vivo* reporter assay to test for transcriptional and translational fidelity in yeast *J. Biol. Chem.* **277** 24420–6

[11] Lynch M 2010 Evolution of the mutation rate *Trends Genet.* **26** 345–52

[12] Gout J-F, Thomas W K, Smith Z, Okamoto K and Lynch M 2013 Large-scale detection of *in vivo* transcription errors *Proc. Natl Acad. Sci. USA* **110** 18584–9

[13] Imashimizu M, Oshima T, Lubkowska L and Kashlev M 2013 Direct assessment of transcription fidelity by high-resolution RNA sequencing *Nucleic Acids Res.* **41** 9090–104

[14] Gout J F *et al* 2017 The landscape of transcription errors in eukaryotic cells *Sci. Adv.* **3** e1701484

[15] Pelechano V, Chávez S and Pérez-Ortín J E 2010 A complete set of nascent transcription rates for yeast genes *PLoS One* **5** e15442

[16] Islam S, Kjällquist U, Moliner A, Zajac P, Fan J-B, Lönnerberg P and Linnarsson S 2011 Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq *Genome Res.* **21** 1160–7

[17] Bai L, Santangelo T J and Wang M D 2006 Single-molecule analysis of RNA polymerase transcription *Annu. Rev. Biophys. Biomol. Struct.* **35** 343–60

[18] Sydow J F and Cramer P 2009 RNA polymerase fidelity and transcriptional proofreading *Curr. Opin. Struct. Biol.* **19** 732–9

[19] Krummel B and Chamberlin M J 1992 Structural analysis of ternary complexes of *Escherichia coli* RNA polymerase *J. Mol. Biol.* **225** 239–50

[20] Nudler E, Kashlev M, Nikiforov V and Goldfarb A 1995 Coupling between transcription termination and RNA polymerase inchworming *Cell* **81** 351–7

[21] Nudler E, Mustaev A, Goldfarb A and Lukhtanov E 1997 The RNA–DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase *Cell* **89** 33–41

[22] Komissarova N and Kashlev M 1997 RNA polymerase switches between inactivated and activated states by translocating back and forth along the DNA and the RNA *J. Biol. Chem.* **272** 15329–38

[23] Shaevitz J W, Abbondanzieri E A, Landick R and Block S M 2003 Backtracking by single RNA polymerase molecules observed at near-base-pair resolution *Nature* **426** 684

[24] Komissarova N and Kashlev M 1997 Transcriptional arrest: Escherichia coli RNA polymerase translocates backward, leaving the 3′ end of the RNA intact and extruded *Proc. Natl Acad. Sci. USA* **94** 1755–60

[25] Galburt E A, Grill S W, Wiedmann A, Lubkowska L, Choy J, Nogales E, Kashlev M and Bustamante C 2007 Backtracking determines the force sensitivity of RNAP II in a factor-dependent manner *Nature* **446** 820

[26] Depken M, Galburt E A and Grill S W 2009 The origin of short transcriptional pauses *Biophys. J.* **96** 2189–93

[27] Hodges C, Bintu L, Lubkowska L, Kashlev M and Bustamante C 2009 Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II *Science* **325** 626–8

[28] Kuhn C-D *et al* 2007 Functional architecture of RNA polymerase I *Cell* **131** 1260–72

[29] Chédin S, Riva M, Schultz P, Sentenac A and Carles C 1998 The RNA cleavage activity of RNA polymerase III is mediated by an essential TFIIS-like subunit and is important for transcription termination *Genes Dev.* **12** 3857–71

[30] Orlova M, Newlands J, Das A, Goldfarb A and Borukhov S 1995 Intrinsic transcript cleavage activity of RNA polymerase *Proc. Natl Acad. Sci. USA* **92** 4596–600

[31] Reinberg D and Roeder R G 1987 Factors involved in specific transcription by mammalian RNA polymerase: II. Transcription factor IIS stimulates elongation of RNA chains *J. Biol. Chem.* **262** 3331–7

[32] Izban M G and Luse D S 1992 The RNA polymerase II ternary complex cleaves the nascent transcript in a 3′→5′ direction in the presence of elongation factor SII *Genes Dev.* **6** 1342–56

[33] Borukhov S, Sagitov V and Goldfarb A 1993 Transcript cleavage factors from *E. coli Cell* **72** 459–66

[34] Voliotis M, Cohen N, Molina-París C and Liverpool T B 2008 Fluctuations, pauses, and backtracking in DNA transcription *Biophys. J.* **94** 334–48

[35] Tripathi T and Chowdhury D 2008 Interacting RNA polymerase motors on a DNA track: effects of traffic congestion and intrinsic noise on RNA synthesis *Phys. Rev.* E **77** 011921

[36] Klindziuk A, Meadowcroft B and Kolomeisky A B 2020 A mechanochemical model of transcriptional bursting *BioPhys. J.* **118** 1213–20

[37] Roldán É, Lisica A, Sánchez-Taltavull D and Grill S W 2016 Stochastic resetting in backtrack recovery by RNA polymerases *Phys. Rev.* E **93** 062411

[38] Sahoo M and Klumpp S 2013 Backtracking dynamics of RNA polymerase: pausing and error correction *J. Phys.: Condens. Matter* **25** 374104

[39] Epshtein V and Nudler E 2003 Cooperation between RNA polymerase molecules in transcription elongation *Science* **300** 801–5

[40] Epshtein V, Toulmé F, Rahmouni A R, Borukhov S and Nudler E 2003 Transcription through the roadblocks: the role of RNA polymerase cooperation *EMBO J.* **22** 4719–27

[41] Jin J, Bai L, Johnson D S, Fulbright R M, Kireeva M L, Kashlev M and Wang M D 2010 Synergistic action of RNA polymerases in overcoming the nucleosomal barrier *Nat. Struct. Mol. Biol.* **17** 745

[42] Nudler E 2012 RNA polymerase backtracking in gene regulation and genome instability *Cell* **149** 1438–45

[43] Galburt E A, Parrondo J M R and Grill S W 2011 RNA polymerase pushing *Biophys. Chem.* **157** 43–7

[44] Liu L F and Wang J C 1987 Supercoiling of the DNA template during transcription *Proc. Natl Acad. Sci. USA* **84** 7024–7

[45] Chong S, Chen C, Ge H and Xie X S 2014 Mechanism of transcriptional bursting in bacteria *Cell* **158** 314–26

[46] Jing X, Loskot P and Yu J 2018 How does supercoiling regulation on a battery of RNA polymerases impact on bacterial transcription bursting? *Phys. Biol.* **15** 066007

[47] Tripathi S, Brahmachari S, Onuchic J N and Levine H 2021 DNA supercoiling-mediated collective behavior of co-transcribing RNA polymerases *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkab1252

[48] Wang H 2009 Stokes efficiency of molecular motors with inertia *Appl. Math. Lett.* **22** 79–83

[49] Kim S, Beltran B, Irnov I and Jacobs-Wagner C 2019 Long-distance cooperative and antagonistic RNA polymerase dynamics via DNA supercoiling *Cell* **179** 106–19

[50] Heberling T, Davis L, Gedeon J, Morgan C and Gedeon T 2016 A mechanistic model for cooperative behavior of co-transcribing RNA polymerases *PLoS Comput. Biol.* **12** e1005069

[51] Belitsky V and Schütz G M 2019 RNA Polymerase interactions and elongation rate *J. Theor. Biol.* **462** 370–80

[52] Toulokhonov I, Zhang J, Palangat M and Landick R 2007 A central role of the RNA polymerase trigger loop in active-site rearrangement during transcriptional pausing *Mol. Cell* **27** 406–19

[53] Chou T and D'Orsogna M R 2014 First passage problems in biology *First Passage Problems in Biology First-Passage Phenomena and Their Applications* ed R Metzler, G Oshanin and S Redner (Singapore: World Scientific) ch 13 pp 306–45

[54] Li W and Lynch M 2020 *eLife* **9** e54898

[55] Traverse C C and Ochman H 2016 Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles *Proc. Natl Acad. Sci. USA* **113** 3311–6

[56] Stevenson-Jones F, Woodgate J, Castro-Roa D and Zenkin N 2020 Ribosome reactivates transcription by physically pushing RNA polymerase out of transcription arrest *Proc. Natl Acad. Sci. USA* **117** 8462–7

[57] Hill T L 1988 Interrelations between random walks on diagrams (graphs) with and without cycles *Proc. Natl Acad. Sci. USA* **85** 2879–83