Bayesian inference and comparison of stochastic transcription elongation models

Jordan Douglas[1,2], Richard Kingston[1], Alexei J. Drummond[2*]

1 School of Biological Sciences, University of Auckland, Auckland, New Zealand
2 Centre for Computational Evolution, Department of Computer Science, University of Auckland, Auckland, New Zealand

* alexei@cs.auckland.ac.nz

Abstract

Transcription elongation can be modelled as a three step process, involving polymerase translocation, NTP binding, and nucleotide incorporation into the nascent mRNA. This cycle of events can be simulated at the single-molecule level as a continuous-time Markov process using parameters derived from single-molecule experiments. Previously developed models differ in the way they are parameterised, and in their incorporation of partial equilibrium approximations.

We have formulated a hierarchical network comprised of 12 sequence-dependent transcription elongation models. The simplest model has two parameters and assumes that both translocation and NTP binding can be modelled as equilibrium processes. The most complex model has six parameters makes no partial equilibrium assumptions. We systematically compared the ability of these models to explain published force-velocity data, using approximate Bayesian computation. This analysis was performed using data for the RNA polymerases of *E. coli*, *S. cerevisiae* and Bacteriophage T7.

Our analysis indicates that the polymerases differ significantly in their translocation rates, with the rates in T7 pol being fast compared to *E. coli* RNAP and *S. cerevisiae* pol II. Different models are applicable in different cases. We also show that all three RNA polymerases have an energetic preference for the posttranslocated state over the pretranslocated state. A Bayesian inference and model selection framework, like the one presented in this publication, should be routinely applicable to the interrogation of single-molecule datasets.

Author summary

Transcription is a critical biological process which occurs in all living organisms. It involves copying the organism's genetic material into messenger RNA (mRNA) which directs protein synthesis on the ribosome. Transcription is performed by RNA polymerases which have been extensively studied using both ensemble and single-molecule techniques (see reviews: [1, 2]). Single-molecule data provides unique insights into the molecular behaviour of RNA polymerase. Transcription at the single-molecule level can be computationally simulated as a continuous-time Markov process and the model outputs compared with experimental data. In this study we use Bayesian techniques to perform a systematic comparison of 12 stochastic models of transcriptional elongation. We demonstrate how equilibrium approximations can strengthen or weaken the model, and show how Bayesian techniques can identify

necessary or unnecessary model parameters. We describe publicly accessible and open-source software that can a) simulate, b) perform inference on, and c) compare models of transcription elongation.

Introduction

Transcription is carried out by RNA polymerases: RNAP in *Escherichia coli*, pol II in *Saccharomyces cerevisiae*, and T7 pol in Bacteriophage T7. It involves the copying of template double-stranded DNA (dsDNA) into single-stranded messenger RNA (mRNA). The template is read in the $3'$ to $5'$ direction, while the mRNA is extended sequentially in the $5'$ to $3'$ direction. RNAP and pol II are comprised of multiple subunits, and their catalytic subunits are homologous [3, 4]. In contrast, T7 pol exists as a monomer with a distinct sequence, and resembles the *E. coli* DNA polymerase I [5].

Optical trapping experiments have been performed on the transcription elongation complex (TEC) from a variety of organisms [6–12]. In a typical experimental setup [6], two polystyrene beads (around 600 nm in diameter) are tethered to the system; one attached to the RNA polymerase and the other to the template DNA. As transcription elongation progresses, the distance between the two beads increases and the velocity of a single TEC can be computed. Optical tweezers can be used to apply a force F to the system. An assisting load ($F > 0$ pN) acts in the direction of elongation while a hindering load ($F < 0$ pN) acts in the opposite direction (Fig 1).

Fig 1. Effect of an applied force on elongation velocity. (A) Optical trapping setup [6] showing dsDNA being transcribed by RNA polymerase (grey ellipse) into mRNA. Two polystyrene beads are tethered to the system allowing forces to be applied using optical tweezers. An assisting load $F > 0$ acts in the same direction as transcription (top) while a hindering load $F < 0$ opposes (bottom). Figure not to scale (the bead has a diameter of 600 nm while the *E. coli* RNAP has an effective diameter of 3.3 nm). (B) The modelled effect of applying a force on RNA polymerase. Five simulations were performed at each force setting for the first 200 nucleotides of the *rpoB* gene. See main text for description of the underpinning transcription elongation model and simulation method.

Such single-molecule studies of the TEC have revealed that RNA polymerases progress in a discontinuous fashion [6, 13–16] with step sizes that correspond to the dimensions of a single nucleotide (3.4 Å [17]). Consequently, at the single molecule level, transcription is best modelled as a discrete process rather than a continuous one.

A single cycle in the main transcription elongation pathway (Fig 2) requires (1) Forward translocation of the RNA polymerase, making the active site accessible; (2) Binding of the complementary nucleoside triphosphate (NTP); (3) Addition of the nucleotide onto the $3'$ end of the mRNA. This third step involves NTP hydrolysis, with a nucleoside monophosphate added onto the chain and pyrophosphate released from the enzyme.

Our study aimed to identify the best model to describe this reaction cycle for RNAP, pol II and T7 pol, based on analysis of published force-velocity data. As there are three reactions, up to six rate constants may be necessary for a kinetic model of a single nucleotide addition. These describe forward and backwards translocation (k_{fwd} and k_{bck}), binding and release of NTP (k_{bind} and k_{rel}), and NTP catalysis and reverse-catalysis (k_{cat} and k_{rev}), also known as pyrophosphorolysis [18]. However fewer than six parameters may be required in practise.

First, it is reasonable to assume that polymerisation is effectively irreversible [19–22], as pyrophosphorolysis is a highly exergonic reaction, reducing the number of rate

Fig 2. State diagrams of RNA polymerase. (A) The model of the main transcription elongation pathway, which shows the postulated states; the pathways for interconversion; and the rate constants that govern each part of the reaction. The transcription bubble is the set of $\beta_1 + h + \beta_2$ bases in the double stranded DNA which are melted. State description denoted by $S(l,t)$ where l is the length of the mRNA and t is the position of the polymerase active site (small grey square) with respect to the $3'$ end of the mRNA. Polymerase translocation displaces the polymerase by a distance of $\delta = 1$ bp $= 3.4$ Å. During polymerisation the chain is extended by one nucleotide and this process could be reversed. (B) Instantiated posttranslocated state of RNA polymerase transcribing the *rpoB* gene sequence, with $\beta_1 = 2, h = 9, \beta_2 = 1$. Translocating forward by a single nucleotide would require melting the $5'$ TA/AT and $5'$ AG/TC doublets (right arrow). Backwards translocation would require melting the two $5'$ AC/TG doublets (left arrow) as well as reconfiguring the mRNA secondary structure.

constants to five. Second, translocation between the pretranslocated and posttranslocated states, and/or NTP binding, may occur on timescales significantly more rapid than the other steps, in which case they may be modelled as equilibrium processes. These assumptions simplify the model, as the respective forward and reverse reaction rate constants can be subsumed by a single equilibrium constant. Third, thermodynamic models of nucleic acid structure can be used to estimate sequence-dependent translocation rates $k_{fwd}(l)$ and $k_{bck}(l)$, and this can sometimes result in parameter reduction [21–23].

If there are L bases in the template then there are in the order of $2L$ translocation rates to estimate to model the main pathway. These rates can be calculated using thermodynamic parameters of nucleic acid basepairing. To translocate forward, two basepairs must be disrupted: (1) the basepair at the downstream edge of the transcription bubble, and (2) the basepair at the upstream end of the DNA/mRNA hybrid Fig 2B. To translocate backwards two different basepairs must be disrupted: (1) the basepair at the upstream edge of the transcription bubble, and (2) the basepair at the downstream end of the DNA/mRNA hybrid. If the $5'$ end of the mRNA has folded then it too may need to be unfolded to facilitate backwards translocation but this is not a component of our models [21, 23].

The energetics of disruption are dependent on the position within the template. The standard Gibbs free energies involved in duplex formation can be calculated using nearest neighbour models. For example in SantaLucia's DNA/DNA parameters [24] $5'$ TA/AT is the weakest doublet with $\Delta G^{(bp)} = -0.58$ kcal/mol while $5'$ GC/CG is the strongest with $\Delta G^{(bp)} = -2.24$ kcal/mol at 37 °C. A good model of translocation should account for the relative strengths of these basepairs when estimating $k_{fwd}(l)$ and $k_{bck}(l)$, or the ratio between the two, at any given position in the template. By accounting for sequence specificity in this fashion, transcriptional pause sites have been predicted with varying levels of success [8, 22, 23].

Irrespective of equilibrium assumptions and parameterisation, transcription elongation under applied force can be modelled in two fundamentally distinct ways. First, there are the **deterministic** equations which can be used to calculate the mean pause-free elongation velocity $v(F, [\text{NTP}])$ as a function of force F and NTP concentration [NTP]. An example is the 3-parameter model [6]:

$$v(F, [\text{NTP}]) = \frac{k_{cat}}{1 + \frac{K_D}{[\text{NTP}]}(1 + K_\tau e^{-F\delta/k_B T})} \qquad (1)$$

where δ is the distance between adjacent basepairs (3.4 Å), $K_D = \frac{k_{rel}}{k_{bind}}$ is the equilibrium constant of NTP binding, $K_\tau = \frac{k_{bck}}{k_{fwd}}$ is the equilibrium constant of translocation, k_B is the Boltzmann constant, and T is the absolute temperature. Increasingly complex equations may be used as more parameters or states are added to the model [6, 8, 21]. Such equations describe the velocity averaged across an ensemble of molecules. Parameter inference applied to velocity-force-[NTP] experimental data is straightforward and computationally fast when using these equations. However these equations do not describe the distribution of velocity nor do they account for site heterogeneity across the nucleic acid sequence and therefore cannot predict local sequence effects.

Second, there are the **stochastic** models, which can be implemented via simulation using the Gillespie algorithm [25]. The mean velocity can be calculated by averaging velocities over a number of simulations for a given F and [NTP]. This offers not just the mean but a full distribution of velocities and could potentially explain emergent properties unavailable from a deterministic model. Unfortunately, simulating can be very slow and therefore parameter inference can be a problem.

In this study we used a Markov-chain-Monte-Carlo approximate-Bayesian-computation (MCMC-ABC) algorithm [26] to estimate transcription elongation parameters for **stochastic** models. The observed pause-free velocities we are fitting to were measured at varying applied force and NTP concentration. For each RNA polymerase under study - *E. coli* RNAP, *S. cerevisiae* pol II, and T7 pol - we fit to one respective dataset from the single-molecule literature [6, 27, 28].

Results

Notation and state space

Suppose the TEC is transcribing a gene of length L. Then let $S(l, t)$ denote a TEC state, where the mRNA is currently of length $l \leq L$, and $t \in \mathbb{Z}$ describes the position of the active site with respect to the $3'$ end of the mRNA. When $t = 0$ the polymerase is pretranslocated and cannot bind NTP, and when $t = 1$ the polymerase is posttranslocated and *can* bind NTP (Fig 2). This study is focused on the main elongation pathway and the observed velocities being fitted have pauses filtered out. Therefore, although additional backtracked states ($t < 0$) [6, 29, 30] and hypertranslocated states ($t > 1$) [31, 32] exist, these are not incorporated in the model.

Parameterisation of the translocation step

While inferences about the rate constants associated with NTP binding and catalysis (k_{bind}, k_{rel}, and k_{cat}) can be made directly from the data, the translocation step is more complex. Chemical rate theory is invoked in order to estimate k_{fwd} and k_{bck}. Recasting the problem in this way (1) allows the sequence-dependence of translocation to be incorporated by considering the energetics of basepairing, and (2) provides a way of accommodating the effects of applied force on the elongation process.

The rates of translocation are calculated from three parameters - $\Delta G_{\tau 1}$, δ_1, and $\Delta G_\tau^\dagger$ - together with the standard Gibbs energies of the pre and posttranslocated states $\Delta G_{S(l,0)}^{(bp)}$ and $\Delta G_{S(l,1)}^{(bp)}$ respectively. Here and elsewhere the superscript (bp) indicates that a term can be evaluated using well-established thermodynamic models of the basepairing energies (see Materials and Methods). The rates $k_{fwd}(l) \equiv k_{fwd}(l|\Delta G_{\tau 1}, \delta_1, \Delta G_\tau^\dagger)$ and $k_{bck}(l) \equiv k_{bck}(l|\Delta G_{\tau 1}, \delta_1, \Delta G_\tau^\dagger)$ are therefore dependent on the local sequence. The physical meaning of the parameters and they way they are used in the model is detailed below.

Energetic bias for the posttranslocated states

$\Delta G_{\tau 1}$ (units $k_B T$) is an additive term used in calculating the standard Gibbs energy of the posttranslocated state. If $\Delta G_{\tau 1} = 0$, then the sequence alone determines the Gibbs energy difference between pre and posttranslocated states. In this case, pretranslocated states are usually favoured over posttranslocated states due to the loss of a single basepair in the hybrid of the latter. If the system has time to reach equilibrium, the probability of observing the pretranslocated state $S(l,0)$ and posttranslocated state $S(l,1)$ are:

$$p(S(l,0)) \propto e^{-(\Delta G^{(bp)}_{S(l,0)})} \tag{2}$$

$$p(S(l,1)) \propto e^{-(\Delta G_{\tau 1} + \Delta G^{(bp)}_{S(l,1)})}. \tag{3}$$

This is described by equilibrium constant K_τ:

$$K_\tau(l) \equiv \frac{p(S(l,0))}{p(S(l,1))}. \tag{4}$$

$\Delta G_{\tau 1}$ has frequently been estimated for T7 pol [33–35] and there has been some discussion around whether such a term is necessary for RNAP [8].

Polymerase displacement and formation of the transition state

δ_1 (units Å) is the distance that the polymerase must translocate forward to facilitate the formation of the transition state. The distance between adjacent basepairs is held constant at an experimentally measured value $\delta = 3.4$ Å [17], and $0 < \delta_1 < \delta$. The response of the system to an applied force F depends on this term. In general, the application of force F tilts the Gibbs energy landscape - the Gibbs energy difference between adjacent translocation states being augmented by a factor $\frac{F\delta}{k_B T}$ (Fig 3 [1,36]).

Fig 3. Effects of translocation parameters on state energies. Sequence independent Gibbs energy landscape of translocation with backtracked states included (for visualisation purposes). The solid red lines represent translocation states ($t = 0$: pretranslocated, $t = 1$: posttranslocated, and $t < 0$: backstepped/backtracked), while the dashed red lines represent transition states. In a translocation equilibrium model the transition states are irrelevant. Applying an assisting force $F > 0$ tilts the landscape in favour of higher values of t. The effect of $\Delta G_{\tau 1}$ is observed at the posttranslocated state $t = 1$.

It may be necessary to estimate δ_1 to model the data adequately [21], or it may be sufficient to simply set $\delta_1 = \delta/2$ [36].

Energy barrier of translocation

$\Delta G^{\dagger}_{\tau}$ (units $k_B T$) is an additive term used in calculating the activation barrier height in the translocation step. Let $T(l,t)$ be the translocation transition state between $S(l,t)$ and $S(l,t+1)$. Then $\Delta G^{\dagger}_{T(l,t)}$ is the sequence-dependent standard Gibbs energy of activation which must be overcome in order to make this translocation. It is assumed this can be written as:

$$\Delta G^{\dagger}_{T(l,t)} = \Delta G^{\dagger}_{\tau} + \Delta G^{(bp)}_{T(l,t)} \tag{5}$$

where $\Delta G^{(bp)}_{T(l,t)}$ is the estimated Gibbs energy of basepairing in the transition state. This term can be sequence-dependent and describes the basepairing energy of the translocation transition state. Bai et al. 2004 [22] implicitly set the value of this term to 0, and therefore the transition state Gibbs energy is determined solely by $\Delta G^{\dagger}_{\tau}$. Tadigotla et al. 2006 [23] estimated $\Delta G^{(bp)}_{T(l,t)}$ as the mean of the basepairing energies of two neighbouring translocation states. We have taken a third approach, where the basepairing configuration of the transition state is approximated by examining the basepairs in the two adjacent translocation states, and the Gibbs energy of the transition state is evaluated directly. While these models could give different predictions of local sequence effects, they are not expected to affect the mean elongation velocity over a long sequence. For more information see S3 Appendix.

Given an applied force F, the translocation rate from the pre to the posttranslocated state $k_{fwd}(l)$ and vice versa and $k_{bck}(l)$ are calculated from this barrier height using the using an Arrhenius type relation:

$$k_{fwd}(l) = Ae^{-(\Delta G^{\dagger}_{T(l,0)} - \Delta G^{(bp)}_{S(l,0)} - F\delta_1/k_BT)} \tag{6}$$

$$k_{bck}(l) = Ae^{-(\Delta G^{\dagger}_{T(l,0)} - (\Delta G^{(bp)}_{S(l,1)} + \Delta G_{\tau 1}) + F(\delta - \delta_1)/k_BT)} \tag{7}$$

where pre-exponential factor A is held constant at 10^6 s^{-1}. This parameter has been arbitrarily set to a variety of values in previous studies ($10^6 - 10^9$ s^{-1} [21–23]). This has little consequence for model fitting, however the value of $\Delta G^{\dagger}_{\tau}$ is entangled with the value of the pre-exponential factor A and can only be meaningfully interpreted in light of this value.

Model space

The full transcription elongation model makes use of the following 6 parameters:

- k_{cat} (units s^{-1}).
- $K_D \equiv \frac{k_{rel}}{k_{bind}}$ (units μM).
- k_{bind} (units μM^{-1} s^{-1}).
- $\Delta G_{\tau 1}$ (units k_BT).
- δ_1 (units Å).
- $\Delta G^{\dagger}_{\tau}$ (units k_BT).

However fewer than 6 parameters may be needed to adequately describe the data. If it is assumed that the energy differences between pre and posttranslocated states are determined by basepairing energies alone, the parameter $\Delta G_{\tau 1}$ does not need to be estimated. If it is assumed that the displacement required for formation of the translocation intermediate state is half the distance between adjacent basepairs, the parameter δ_1 does not need to be estimated. Partial equilibrium approximations may also simplify the model. If binding is approximated as an equilibrium process, k_{bind} does not need to be estimated. If translocation is approximated as an equilibrium process, $\Delta G^{\dagger}_{\tau}$ and δ_1 do not need to be estimated. One, both, or neither of these two steps could be assumed to achieve equilibrium, thus yielding four equilibrium model variants (Fig 4A).

NTP binding has been modelled as both a kinetic and equilibrium process in the literature [6, 21, 22]. In a kinetic binding model, NTP binding occurs at pseudo-first

Fig 4. The space of models to be compared. (A) The four equilibrium model variants. NTP binding, translocation, both, or neither, could be assumed to achieve equilibrium prior to catalysis. (B) The 12 transcription elongation models. An arrow connects model i to j if and only if a single parameter can be included with model i to obtain model j. Number of parameters to estimate k is shown for each level in the network. Equilibrium approximation colour scheme the same as in A. $\Delta G_{\tau 1}$ and δ_1 can each be estimated or set to a constant.

order rate k_{bind}[NTP], while NTP release occurs at rate k_{rel}, and k_{bind} and $\frac{k_{rel}}{k_{bind}}$ are the parameters to estimate. Under a partial equilibrium approximation NTP binding and release are assumed to be rapid enough that equilibrium is achieved. In this case, the rate constants k_{bind} and k_{rel} are subsumed by the NTP dissociation constant $K_D \equiv \frac{k_{rel}}{k_{bind}}$ which becomes the sole binding-related parameter to estimate. The introduction of partial equilibrium approximations for both the NTP binding and translocation steps has implications when specifying the prior distributions for the Bayesian analysis (S4 Appendix.)

Incorporating these simplifications to the model in a combinatorial fashion results in a total of 12 related models, which together constitute the model space. Our objective was to determine which of these 12 models provides the best description of the experimental data. The simplest model (Model 1) contains 2 parameters (k_{cat} and K_D). The most complex model (Model 12) contains all 6 parameters. The full model space is displayed in Fig 4B.

Stochastic modelling

For each model we performed stochastic simulations, appropriate for the modelling of single-molecule force-velocity data. The simulations, performed using the Gillespie algorithm [25, 37], can be used to estimate the mean elongation velocity under a model.

The estimation of mean velocity can be broken down into three steps. First, the system is initialised by placing the RNA polymerase at the $3'$ end of the template, with the transcription bubble open and a DNA/RNA hybrid formed. The force and NTP concentrations are assigned their experimentally set values. Second, a chemical reaction is randomly sampled. The probability that reaction $S \xrightarrow{k} S'$ is selected is proportional to its rate constant k (Fig 2). The amount of time taken for the reaction to occur is sampled from the exponential distribution. States which are subject to a partial equilibrium approximation are coalesced into a single state, which augments the outbound rate constants. The second step is repeated until the RNA polymerase has copied the entire template. Third, the previous two steps are repeated c times. The mean elongation velocity is evaluated as the mean of each mean elongation velocity across c simulations. For more information, see S1 Appendix.

Model selection with MCMC-ABC

Our aim was to 1) use Bayesian inference to select the best of 12 models for each RNA polymerase; and 2) estimate the parameters for those of the 12 transcription elongation models which appear in the 95% credible set of the posterior distribution. Selecting prior distributions behind each parameter is a critical process in Bayesian inference. A prior distribution should reflect what is known about the parameter before witnessing the new data. We have explicated our prior assumptions, with justifications, in Table 2.

We performed MCMC-ABC experiments which estimated the parameters and model indicator M_i for $i \in \mathbb{Z}, 1 \leq i \leq 12$. Models which appear more often in this posterior distribution are better choices, given the data. It is necessary to achieve a large effective sample size (ESS) by running the MCMC chains sufficiently long in order to provide accurate parameter estimates for all sampled models.

Based on the MCMC-ABC experiments (Table 1) the best models for the datasets examined are Models 11 and 12 for both RNAP and pol II, and Model 5 for T7 pol.

For the *S. cerevisiae* pol II, Model 12 has the highest posterior probability $P(M_{12}|D) = 0.783$. This is the most complex model considered, with 6 estimated parameters. In Model 12 translocation, NTP binding and catalysis are all kinetic processes; the displacement required to facilitate formation of the translocation transition state, $\delta_1 < \delta$, is estimated ($\hat{\delta_1} = 3.124$ Å); and the standard Gibbs energy of the posttranslocated state is influenced by parameter $\Delta G_{\tau 1} \neq 0$.

The posterior distribution for the *E. coli* RNAP consists of the same set models as that of *S. cerevisiae* pol II. For RNAP, Model 11 has the highest probability $P(M_{11}|D) = 0.863$. This model is a submodel of Model 12 with one fewer parameter: in Model 11 NTP binding is treated as an equilibrium process while in Model 12 it is not.

The only model in the 95 % credible set for the RNA polymerase of Bacteriophage T7 is Model 5 $P(M_5|D) = 0.993$. In Model 5 (4 parameters) translocation, but not binding, is treated as an equilibrium process, and $\Delta G_{\tau 1}$ is estimated. This positions T7 pol in a quite different area of the model space to the other two polymerases. Plots showing how the models fit all three datasets are displayed in S3 Fig. These plots show that the T7 pol model does not explain the data as well as that of RNAP or pol II.

A high ESS (> 100-200 [38]) is essential for making reliable parameter estimates. Almost all parameters in the 95 % credible set of models are sufficiently estimated by this criterion. Unfortunately, the ESS of $\frac{k_{rel}}{k_{bind}}$ is very low in the kinetic binding models (Model 5 for T7 pol and Model 12 for RNAP/pol II), particularly for RNAP where the ESS = 11. In contrast to the other parameters, obtaining a sufficient ESS for $\frac{k_{rel}}{k_{bind}}$ was not computationally realisable. The fundamental issue, apparent from simulations (see S4 Appendix) is that the datasets do not inform on NTP binding kinetics.

Discussion

In this paper we evaluated some simple Brownian ratchet models of transcription elongation (Fig 2). By varying the parameterisation of the translocation step (Fig 3) and incorporating partial equilibrium approximations commonly invoked in the literature (Fig 4A) we enumerated a total of 12 related models (Fig 4B). Using stochastic simulations and approximate Bayesian computation, we then assessed which of these models were capable of describing the force-velocity data previously measured for several RNA polymerases using single-molecule optical trapping experiments [6, 27, 28].

Our analysis suggests that 1) different partial equilibrium approximations are appropriate for the multisubunit RNA polymerases versus the single subunit T7 RNA polymerase. 2) Treatment of the NTP binding step remains a point of ambiguity. The existing data does not place strong constraints on the modelling of this step. 3) There is an energetic bias for posttranslocated state. 4) The model of the translocation step, which evokes chemical rate theory, is not physically realistic.

Relation to previous models and stochastic simulations

There is an extensive literature concerned with the kinetic modelling of transcription elongation. Such models may incorporate backtracking, hypertranslocation, and other reactions. Here we are concerned with only with the central elongation pathway as we fit the models to pause-edited single-molecule data.

Table 1. Summary of posterior distributions from MCMC-ABC experiments.

Enzyme	*E. coli* RNAP		*S. cerevisiae* pol II		Bacteriophage T7 pol
i Description of M_i	11 Binding equilibrium, Translocation kinetic	12 Binding kinetic, Translocation kinetic	11 Binding equilibrium, Translocation kinetic	12 Binding kinetic, Translocation kinetic	5 Binding kinetic, Translocation equilibrium
$\hat{k}_{cat}$ (s^{-1})	25.93 (25.2, 26.68)	26.44 (25.21, 26.73)	28.63 (27.6, 29.84)	28.63 (27.56, 29.88)	127.4 (124.1, 131.2)
$\widehat{\frac{k_{rel}}{k_{bind}}}$ (μM)	97.84 (90.74, 104.9)	99.04 (25.76, 98.07)	75.32 (68.51, 82.89)	11.78 (7.179E-3, 57.27)	22.07 (9.521E-3, 65.69)
$\hat{k}_{bind}$ (μM^{-1} s^{-1})	–	6.779 (0.3225, 9.489)	–	0.4635 (0.346, 1.619)	1.519 (1.082, 3.072)
$\Delta\hat{G}_{\tau 1}$ ($k_B T$)	-1.925 (-2.08, -1.819)	-2.022 (-2.071, -1.802)	-4.667 (-5.762, -3.641)	-4.687 (-5.797, -3.624)	-4.048 (-4.232, -3.776)
$\Delta\hat{G}_{\tau}^{\dagger}$ ($k_B T$)	5.003 (4.922, 5.113)	5.072 (4.917, 5.109)	4.585 (4.466, 4.747)	4.601 (4.451, 4.746)	–
$\hat{\delta}_1$ (Å)	3.246 (3.028, 3.4)	3.147 (3.028, 3.4)	3.092 (2.778, 3.399)	3.124 (2.773, 3.4)	–
ESS	343 (322, 462)	118 (11, 212)	331 (228, 599)	816 (79, 1730)	516 (26, 1775)
X^2	2.37	2.37	0.693	0.666	4.613
$P(M_i \mid D)$	0.863	0.137	0.217	0.783	0.993

Each column summarises the posterior distribution for the respective RNA polymerase, which arises from multiple independent MCMC chains. These estimates are conditional on the model M_i specified at the top of the column, and only models which appear in an RNA polymerase's 95% credible set have estimates. Parameter estimates ($\hat{\theta}$; geometric median) and 95% HPD shown. A '–' is left in place if the parameter is not included in the model. Effective sample sizes (ESS) and HPD intervals calculated with Tracer 1.6 [38]. The ESS displayed is the mean ESS across all parameters estimated for that model (lowest ESS, highest ESS). A parameter should have an ESS greater than 100-200 in order to have a reliable estimate. The chi-squared test statistic X^2 of the geometric median and posterior probability $P(M_i|D)$ are shown for each model.

A stochastic and sequence-dependent model was proposed by Bai et al. 2004 [22] for RNAP, with both NTP binding and translocation treated as equilibrium processes. The translocation equilibrium constant was calculated entirely from basepairing energies. Therefore this model is equivalent to Model 1, and the parameters were estimated as $k_{cat} = 24.7$ s^{-1} and $K_D = 15.6$ μM from fit to experimental data. Maoiléidigh et al. 2011 also presented stochastic simulations of RNAP. The elongation component of their model is equivalent to either models 1, 2, 3, or 6, dependent on the choice of parameters (the parameters are presented in Table S3 of [21]). We build on this work by providing a systematic Bayesian framework for model comparison and parameter estimation.

While our analysis employed sequence-dependent stochastic models, comparisons can also be made with some deterministic models.

Abbondanzieri et al. 2005 [6], Larson et al. 2012 [39] , Schweikhard et al. 2014 [27], and Thomen et al. 2008. [28,40] described a deterministic model (for RNAP, pol II, pol II, and T7 pol respectively) which estimated k_{cat}, K_D and translocation equilibrium constant $K_\tau = \frac{k_{bck}}{k_{fwd}}$. These are most similar to Model 4.

Maoiléidigh et al. 2011 for RNAP, and Dangkulwanich et al. 2013 for pol II, however found that the translocation and catalysis were occurring on similar timescales, and modelled only NTP binding as an equilibrium process [21,35]. They also estimated the distance of translocation. These deterministic models are most similar to Model 11.

Finally, Mejia et al. 2015 [41] used a model that is quite different to all the above models, as it does not explicitly treat translocation. Instead elongation is modelled with a two step kinetic scheme, the first step involving NTP binding and conformational change, and the second step involving nucleotide incorporation and product release. This model is most similar to a special case of Model 5 where $\Delta G_{\tau 1}$ becomes extremely negative, driving the polymerase into the posttranslocated position.

Translocation rates differ among RNA polymerases

For RNAP and pol II, we estimate that a partial equilibrium approximation for the translocation step is inadequate. The posterior probability that such models are inadequate is 1.00 (see Table 1). For T7 pol, however, translocation is significantly faster than catalysis and is best modelled with a partial equilibrium approximation. This same posterior probability is 0.00. Using estimates for $\Delta G_{\tau}^{\dagger}$ and $\Delta G_{\tau 1}$ under the maximum posterior models (Model 11 for RNAP and Model 12 for pol II) we estimate the mean forward $\bar{k}_{fwd}$ and backward $\bar{k}_{bck}$ translocation rates averaged across the *rpoB* sequence as: 268.5 s^{-1} and 116 s^{-1} for RNAP, and 401.3 s^{-1} and 10.96 s^{-1} for pol II, respectively. These estimates are within one order of magnitude of the respective estimate for the rate of catalysis (see Table 1) suggesting that translocation and catalysis indeed occur on similar timescales.

For RNAP and pol II, translocation has frequently been modelled as an equilibrium process [6, 22, 27, 39, 41], however in some recent analyses this assumption has been rejected [21, 23, 35, 42, 43]. Our Bayesian analysis supports this. In contrast, there is general agreement that translocation in T7 pol is adequately modelled as an equilibrium process [28, 44, 45].

The data does not determine the kinetics of the NTP binding step

It remains unclear how to best model the NTP binding step. Models that describe NTP binding as a kinetic process have posterior probabilities of 0.14 for RNAP, 0.78 for pol II and 0.993 for T7 pol (Table 1). However, in our sensitivity analysis, where we used different prior distributions for K_D, these probabilities were 0.65, 0.22, and 0.19, respectively (results not shown). The sensitivity of the probabilities to the choice of prior and their intermediate magnitude, implies the data carry little information about the rates of NTP binding and release

Furthermore, our kinetic-binding-model estimates for $\frac{k_{rel}}{k_{bind}}$ and sometimes k_{bind} have HPD intervals which span multiple orders of magnitude (Table 1), again suggesting that it is not achievable to estimate both of these parameters simultaneously from this data alone.

The pause-free mean velocities measured during transcription elongation follow Michaelis-Menten kinetics [46] even though the reaction cycle is more complicated than that of a simple enzyme. As such, the inability to resolve the timescale of the substrate binding step is unsurprising [47–49].

NTP binding is almost always assumed to achieve equilibrium for RNAP, pol II and T7 pol [6, 21–23, 27, 28, 35, 39, 40, 45]. However Mejia et al. 2015 have suggested that NTP binding is rate-limiting, based on experimental manipulation of the RNAP trigger loop [41] .

RNAP has an energetic preference for the posttranslocated state

In previous stochastic sequence-dependent models [22, 23] the standard Gibbs energies of the pre and posttranslocated states have been based solely on the nucleic acid basepairing energies. Our models include an additional term, $\Delta G_{\tau 1}$, to account for potential interactions between the protein and the nucleic acid. The marginal posterior probability of a model in which an additional term $\Delta G_{\tau 1}$ is required is 1.00 in all three polymerases. In each case $\Delta G_{\tau 1}$ was estimated to be less than 0 $k_B T$ and 0 $k_B T$ is not included in the 95% highest posterior density (HPD) (Table 1). We find that $\Delta \hat{G_{\tau 1}}$ is the most significant in pol II and T7 pol: -4.687 $k_B T$ and -4.048 $k_B T$ respectively, while $\Delta \hat{G_{\tau 1}} = -1.925$ $k_B T$ for RNAP.

These results suggest that structural elements within RNA polymerases can energetically favour posttranslocated states over pretranslocated states. We note that the sequence-dependent contribution of the dangling end of the DNA/RNA hybrid (estimated to stabilise the posttranslocated state by about -0.8 $k_B T$ at room temperature [22, 24]) is included in the thermodynamic model. The energetic bias for the posttranslocated state is separable from this effect.

To facilitate comparison with previous deterministic models, using our estimates of $\Delta G_{\tau 1}$ we calculated the equilibrium constant between the pre and posttranslocated states. Geometrically averaged across the *rpoB* gene, these are

$$\bar{K_\tau} = \frac{1}{L - h - \beta_1 - 1} exp\{\sum_{l=h+\beta_1+2}^{L} \ln(k_{bck}(l)/k_{fwd}(l))\} = \begin{cases} 0.7872 \text{ for RNAP} \\ 0.04972 \text{ for pol II} \\ 0.0942 \text{ for T7 pol.} \end{cases} \quad (8)$$

Thus, for all three polymerases, the small energetic preference that the protein has for the posttranslocated state is sufficient to override the loss of basepairing energy, thereby biasing the system towards population of the posttranslocated positions. This is in agreement with some estimates for T7 pol and pol II which place K_τ less than 1 [27, 28, 33, 34, 39] and KIreeva et al. 2018 [43]: *"forward translocation occurs in milliseconds and is poorly reversible"*. However these estimates are inconsistent with some RNAP and pol II studies which place this ratio above 1 [6, 21, 35].

Kinetic modelling can itself suggest no physical mechanism for the stabilization. Yu et al. 2012 [34] have identified a conserved tyrosine residue near the active site of T7 pol that pushes against the $3'$ end of the mRNA, and thus stabilises the posttranslocated state. They propose a similar mechanism for the multi-subunit RNA polymerases.

δ_1 may be an important parameter but its physical meaning is unclear

Our results suggest that δ_1, the distance that RNA polymerase must translocate forward by to reach the translocation transition state, is a necessary parameter to estimate for RNAP and pol II. Setting $\delta_1 = \delta/2$ is not sufficient. The marginal posterior probability of models which estimate this term is 1.00. δ_1 is irrelevant to the modeling of the T7 pol data because the best models invoke a partial equilibrium approximation for the translocation step.

While our prior distribution restricted δ_1 to lie in the range $(0, \delta)$, the upper end our 95% HPD intervals of δ_1 for RNAP and pol II are very close to $\delta = 3.4$ Å (3.400 Å (4 sf) for both polymerases). If it was not for this prior distribution, δ_1 estimates would have included values higher than δ. Similar results have been observed by Maoiléidigh et al. 2011 [21] for RNAP using the same dataset.

Our interpretation of δ_1 implies it should never be greater than δ nor should δ be more than the width of one basepair. The physical meaning of δ_1 with values greater than δ is thus unclear. It is worth noting that δ_1 is only used when $F \neq 0$.

Comparing the kinetics of RNA polymerases

The *in vivo* rate of transcription elongation varies considerably across RNAP, pol II and T7 pol. The prokaryotic and eukaryotic RNA polymerases have a mean rate ranging from 25-100 bp/s [50–56], which may be slowed down in histone-wrapped regions of eukaryotic genomes [9]. Meanwhile Bacteriophage T7 pol operates an order of magnitude faster (around 200-240 bp/s [50, 57]). Furthermore, T7 pol is known to be quite insensitive to pause sites during transcription elongation [11, 28].

In additional to these differences, we have shown that translocation of T7 pol is very rapid, while translocation of RNAP and pol II is a slower process. Furthermore, the model does not fit the data for T7 pol as closely it does for the other two (S3 Fig). T7 pol therefore seems to operate under quite a different kinetic scheme than that of RNAP and pol II, which is not unexpected given its distant evolutionary relationship with the cellular polymerases [5].

In general the velocity of RNA polymerases is significantly slower in an optical trap (estimates ranging from 9.7-22 bp/s for RNAP [13–15, 41, 58]) than the velocity of the untethered enzyme (estimates *in vitro* or *in vivo* ranging from 25-118 bp/s for RNAP [50, 51, 59, 60]). This relationship holds for multiple RNA polymerases including *E.coli* RNAP, *S.cerevisiae* pol II [35, 39, 61, 62], Bacteriophage T7 pol [11, 28, 50, 63], and Bacteriophage Φ6 P2 [12, 64]. This suggests that optical trapping perturbs the system to a significant extent. Additionally, varying degrees of heterogeneity in elongation rate have been observed across different polymerase complexes under the same conditions [13, 15, 28].

The velocity perturbations resulting from the optical trapping apparatus will be propagated into the model parameters eg. $k_{cat}, k_{bind}, \Delta G_\tau^\dagger$, and some caution is needed when extrapolating these results to untethered systems.

Bayesian inference of transcription elongation

To our knowledge we are the first to perform Bayesian inference on single-molecule models of transcription elongation. This was achieved by simulation which necessitated the use of approximate Bayesian computation. An alternative would be to build and use a likelihood function, which would involve constructing a rate matrix for the transcription elongation pathway and exponentiating the rate matrix to calculate the likelihood (ie. the probability of taking exactly t units of time for RNA polymerase to copy the sequence n times). This would be computationally faster however a multitude of numerical issues can arise from matrix exponentiation when the rate matrix is large, or has an absorbing state. This is an approach which we would like to explore in the future.

We have demonstrated that single-molecule data can be usefully analysed using a Bayesian inference and model selection framework. This analysis would have even greater statistical power if applied to the progression of individual RNA polymerase complexes instead mean velocities obtained from multiple experiments.

Materials and methods

Nucleic acid thermodynamics

To estimate the standard Gibbs energies $\Delta_r G^0$ of basepairing, we used SantaLucia's parameters for DNA/DNA basepairs [24] and Sugimoto's parameters for DNA/RNA basepairs [65]. Gibbs energies are described relative to the thermal energy of the system by expressing $\Delta^r G^0 (= \Delta G)$ in units of $k_B T$ instead of kcal/mol, where $k_B T = 0.6156$ kcal/mol at $T = 310$ K. We set $T = 310$ K as this is the temperature the nearest neighbour models were constructed at [24, 65].

The basepairing Gibbs energy $\Delta_{total}^{(bp)}$ for a given state is the sum of basepairing plus dangling end Gibbs energies in the DNA/RNA hybrid $\Delta_{hybrid}^{(bp)}$ plus the basepairing Gibbs energies in the DNA/DNA gene $\Delta_{gene}^{(bp)}$, ie. $\Delta_{total}^{(bp)} = \Delta_{hybrid}^{(bp)} + \Delta_{gene}^{(bp)}$.

Software and algorithms

The effective diameter of RNAP (Fig 1) was calculated from an RNAP crystal structure (PDB: 4MEY [66]) using 3V [67]. Effective diameter is defined as the diameter of a sphere which has the same surface-area to volume ratio as the original object.

Single-molecule simulations were performed using the Gillespie algorithm [25]. These datasets we fit to are all from the single-molecule literature and are presented in: Figures 5a and 5b of Abbondanzieri et al. 2005 [6] for *E. coli* RNAP, Figure 2a of Schweikhard et al. 2014 [27] for *S.cerevisiae* pol II, and Table 2 of Thomen et al. 2008 [28] for T7 pol. To computationally replicate these experiments as faithfully as we could with the available information and computational limitations, simulations in this study were run on the 4 kb *E. coli rpoB* gene for RNAP (GenBank: EU274658), the first 4.75 kb of the human *rpb1* gene for pol II (NCBI: NG_027747) the first 24 kb of the Enterobacteria phage λ genome for T7 pol (NCBI: NC_001416). The mean velocities from 32 (for RNAP), 10 (for pol II) and 3 (for T7 pol) simulations of the full respective sequences were used to estimate the mean elongation velocity during MCMC-ABC, given F and [NTP].

A hybrid length of $h = 9$ bp, upstream-bubble size of $\beta_1 = 2$ bp and downstream-bubble size of $\beta_2 = 1$ bp were used [21, 68]. While there is some uncertainty in these parameters, and they may differ between RNA polymerases, they are not expected to have any effect on mean elongation velocity over a long sequence and were thus held constant.

We used an MCMC-ABC algorithm for parameter inference and model selection [26, 69]. A heavy-tailed distribution [70] was used as a proposal distribution where the parameter to change is selected uniformly at random. To achieve burn-in we used an exponential cooling scheme on ϵ [71] where $\epsilon_{i+1} = \max(\epsilon_{min}, \epsilon_i \gamma)$ for manually tweaked values of $0 < \gamma < 1$ and ϵ_0. Chains which failed to converge were discarded.

For model selection we ran one or more independent MCMC-ABC chains for each selected ϵ_{min} / RNA polymerase combination. Selecting the threshold ϵ_{min} is a critical process in approximate Bayesian computation. Threshold ϵ_{min} must be large enough to achieve convergence with finite computational resources, but small enough that the resulting posterior distribution is still an accurate approximation of the true posterior distribution. We evaluated model fit using the chi-squared test statistic X^2. This means that only model-parameter samples which simulated a set of elongation velocities that closely agree with the data, such that $X^2 < \epsilon_{min}$, were accepted into the respective posterior distribution. For each RNA polymerase we set ϵ_{min} to some initial guess. Then we ran the MCMC chain until the ESS for X^2 was large (> 300) and lowered ϵ_{min} to the bottom 0.05 quantile of the posterior distribution of X^2. This step was repeated until either: a) the distribution of model indicators M converged (model posterior probabilities have changed by less than 0.01, on average). Or, b) the acceptance rate was less than 5%. This was because when the acceptance rate became low, continuing to lower ϵ_{min} became too expensive. The values of ϵ_{min} we ended up using in the final posterior distributions were 2.39 for RNAP, 0.705 for pol II, and 4.63 for T7 pol.

Effective sample sizes and 95% highest posterior density (HPD) intervals were calculated using Tracer 1.6 [38]. Parameters were estimated using the geometric median: that is the posterior sample which has the minimum average Euclidean distance from all other posterior samples. Parameters were normalised into z-scores first. Our code is open source and available at http://www.polymerase.nz. Textfiles containing the posterior distributions and simulation settings are available to download or visualise with the software.

Prior distributions

Prior distributions are a crucial part of any Bayesian analysis. A prior should reflect one's belief about what values a parameter may take *before* witnessing the new data. In our case the prior distributions must not be derived from the data presented by Abbondanzieri et al. 2005 [6] for RNAP, Schweikhard et al. 2014 [27] for pol II, or Thomen et al. 2008 [28] for T7 pol.

Selecting prior distributions for k_{cat}, $\Delta G_{\tau 1}$ and δ_1 were straightforward and where possible based directly off estimates from the literature. Whereas to select priors for parameters $\Delta G_{\tau}^{\dagger}$, k_{bind} and $\frac{k_{rel}}{k_{bind}}$ we used simulations to establish the range of parameter values we would expect under the model. See S4 Appendix and S2 Fig.

Supporting information

S1 Appendix. Stochastic simulation.

Reactions are simulated using the Gillespie algorithm [25]. Given the current state s and a set of possible reactions $s \to s_1, s \to s_2, \ldots, s \to s_n$ with rate constants $k_1, k_2, \ldots, k_n$, the next reaction to perform is sampled proportional to its rate:

$$p(s \to s_i) = \frac{k_i}{\sum\limits_{j=1}^{n} k_j}. \tag{9}$$

The amount of time the reaction takes to occur is sampled from the exponential distribution with rate $\sum\limits_{j=1}^{n} k_j$.

S2 Appendix. MCMC-ABC. Given model/parameters Θ and observed data $D = (D_1, D_2, \ldots, D_n)$, Bayesian inference conventionally involves approximating the posterior probability distribution $P(\Theta|D)$ using the likelihood $P(D|\Theta)$ and the prior $P(\Theta)$:

$$P(\Theta|D) \propto P(D|\Theta)P(\Theta). \tag{10}$$

We do not have a likelihood function and instead must simulate, using the chi-squared test statistic X^2 to evaluate how well a given set of parameters fits the data:

Table 2. Prior distributions used during Bayesian inference.

Parameter	Prior distribution(s)	Justification of prior distribution(s)
Model M	$P(M_i) = 2/16$ for $i \in \{1, 2, 4, 5\}$ $P(M_i) = 1/16$ for $i \in \{3, 6, 7, 8, 9, 10, 11, 12\}$	The three model settings should each have uniformly distributed values. Models with equilibrium translocation have double the prior probability since these models do not use δ_1.
k_{cat} (s^{-1})	Lognormal($\mu = 3.454$, $\sigma = 0.587$) **for RNAP/pol II** Lognormal($\mu = 4.585$, $\sigma = 0.457$) **for T7 pol**	k_{cat} and elongation velocity estimates for *E. coli* RNAP and *S.cerevisiae* pol II range from 18 to 50 s^{-1} for optical trapping experiments [8–10, 22, 41], while elongation velocity goes up to 100 bp/s *in vivo* [51, 52, 55, 72]. Distribution selected such that $(10, 100)$ is central 95% interval. For T7 pol k_{cat} and elongation velocity estimates range from 43 - 240 [11, 50, 63, 73]. Distribution selected such that $(40, 240)$ is central 95% interval.
K_D (μM)	Lognormal($\mu = 1.844$, $\sigma = 1.762$)	Estimates for K_D under binding equilibrium models range from 20-140 μM [8, 20, 39, 40, 62]. In models where binding is kinetic and slow, $K_D \equiv \frac{k_{rel}}{k_{bind}}$ could be much lower (S4 Appendix). To accomodate for both binding models, the prior distribution was selected such that the central 95% interval is (0.2, 200).
k_{bind} (μM^{-1}s^{-1})	Lognormal($\mu = -1.498$, $S\sigma = 1.585$)	Central 95% interval set so that NTP binding is a slow kinetic step (S4 Appendix). Centered around (0.01, 5).
$\Delta G_{\tau 1}$ ($k_B T$)	Normal($\mu = 0$, $\sigma = 1.55$) **for RNAP/pol II** Normal($\mu = -3.3$, $\sigma = 1.55$) **for T7 pol**	For RNAP and pol II, centered around 0 with a standard deviation comparable to the free energy of a single nucleotide basepair doublet, and such that the 95% central interval is (-4, 4). For T7 pol $\Delta G_{\tau 1}$ has been estimated as -4.3 [40] and -4.87 $k_B T$ [33]. However these estimates are likely resulting partially from dangling ends. Thus, we subtracted the mean dangling end contribution of $\sim$ -1 $k_B T$ [24] and centered the prior around this interval with a standard deviation the same as above.
$\Delta G_{\tau}^{\dagger}$ ($k_B T$)	Normal($\mu = 5.5$, $\sigma = 0.97$) **for RNAP/pol II** Normal($\mu = 2.5$, $\sigma = 1.36$) **for T7 pol**	Central 95% interval set so that translocation is a slow kinetic step (S4 Appendix). Selected so that 99% central interval is (3, 8) for RNAP and pol II, and (-1, 6) for T7 pol.
δ_1 (Å)	Uniform($l = 0$, $u = 3.4$)	Uniformly distributed across all possible values.

Prior distributions behind all estimated parameters and the model indicator. Unless specified otherwise, the prior distribution is used for all three RNA polymerases. Lognormal priors (parameterised in log space) are used for rates and equilibrium constants while normal priors are used for Gibbs energy terms.

$$X^2 = \sum_{i=1}^{n} \frac{(S_i - D_i)^2}{S_i} \tag{11}$$

where S_i is the mean velocity simulated under the same [NTP] and assisting force F

that D_i was measured under. The probability that $X^2 = 0$ is equal to the likelihood $P(D|\Theta)$, however it is computationally impractical to only accept parameters into the posterior distribution when the simulation yields $X^2 = 0$. Therefore a threshold ϵ is used and sample Θ_i is accepted into the posterior only if $X^2(\Theta_i) < \epsilon$. This method is called approximate Bayesian computation [26, 69]. This is coupled with Markov chain Monte Carlo (MCMC) to give the MCMC-ABC algorithm which is becoming increasingly popular among computational biologists [26, 71].

For each RNA polymerase we performed one MCMC which, alongside the parameters, additionally estimated model indicator M. Each elongation model uses up to 6 parameters in its simulation component. This means that the 12 models are sharing the same set of estimates in the MCMC. There are therefore seven terms for this MCMC to estimate: M, k_{cat}, $\frac{k_{rel}}{k_{bind}}$, k_{bind}, $\Delta G_{\tau 1}$, $\Delta G_{\tau}^{\dagger}$, and δ_1. When current model M_i requires a parameter to be held constant (eg. in Model 1 $\Delta G_{\tau 1} = 0$), then this parameter would be set to its constant during the simulation without affecting or being affected by its current MCMC estimate.

S3 Appendix. Approximating the translocation transition state.

We demonstrate four methods that can be used to estimate the Gibbs energy of a translocation transition state:

$$\Delta G_{T(l,t)}^{(bp)} = \begin{cases} 0 & \text{for absolute model} \\ \frac{\Delta G_{S(l,t)}^{(bp)} + \Delta G_{S(l,t+1)}^{(bp)}}{2} & \text{for midpoint model} \\ \Delta G_{S(l,t) \cap S(l,t+1)}^{(bp)} & \text{for melting model} \\ \Delta G_{S(l,t) \cup S(l,t+1)}^{(bp)} & \text{for sealing model.} \end{cases} \tag{12}$$

These translocation models are shown in S1 Fig. The first model, which we refer to as the absolute model, is by Bai et al. 2004 [22]. This assumes that the translocation barrier's absolute height is constant for all positions in the template - the energy needed to transit is only dependent on the energy of the translocation state and the value of $\Delta G_{\tau}^{\dagger}$. We refer to the second model, introduced by Tadigotla et al. 2006 [23], as the midpoint model. This model has the desirable property of $\Delta G_{\tau}^{\dagger}$ being the mean barrier height across the sequence. While these two models are simplistic, they have both met the prediction of pause site positions with some success.

Here we demonstrate two more complex models which could be used; both of which estimate the Gibbs energies of basepairing within the transition state and thus describe physical mechanisms. In the melting model, forward translocation involves: first a basepair on the front of the transcription bubble melts (ie. the bubble grows wider by 1 nt) and a basepair on the back of the hybrid melts. Second, RNA polymerase moves forwards into the new opening. Third, a basepair on the back of the transcription bubble and on the front of the hybrid form. Thus the polymerase has translocated forward by one nucleotide. The transition state between translocation states s and s' is comprised of 1) dsDNA with all the basepairs found in both s and s' (ie. the intersection), and 2) a hybrid with all the basepairs found in both s and s' (the intersection). This TEC has Gibbs energy $\Delta G_{S(l,t) \cap S(l,t+1)}^{(bp)}$.

The sealing model describes a process similar to the melting model however in the reverse order: first a basepair on the back side of the bubble forms (and the bubble seals); second this closing propels RNAP forward by one nucleotide; third a basepair on the front of the bubble melts. This transition state is therefore comprised of 1) dsDNA with all the basepairs found in either s or s' (the union), and 2) a hybrid with all the basepairs found in both s and s' (the intersection). This TEC has Gibbs energy $\Delta G_{S(l,t) \cup S(l,t+1)}^{(bp)}$.

The melting and sealing model describe quite different physical mechanisms of translocation. While these models could give different predictions of local sequence effects, they are not expected to affect the mean elongation velocity over a long sequence. Therefore the experimental data we are evaluating can not discriminate between these four models. For the entirety of this paper we used the melting model. Further investigation will be necessary to determine the accuracy of these four models.

S4 Appendix. Prior distributions.

Prior for $\Delta G_\tau^\dagger$, which governs the rates of translocation

RNAP/pol II: to select a prior for $\Delta G_\tau^\dagger$ we simulated transcription on the *rpoB* gene under Model 3 - the simplest binding equilibrium model. $\Delta G_\tau^\dagger$ and k_{cat} were sampled uniformly from a relevant range, with K_D held constant at 100 μM and [NTP] = 1000 μM. For each simulation, the mean elongation velocity was calculated. The results are displayed in S2 Fig.

This plot shows that as the energy barrier of translocation ($\Delta G_\tau^\dagger$) increases, the velocity decreases. If $\Delta G_\tau^\dagger \gtrsim 8$ $k_B T$ then it becomes impossible to achieve a realistic mean velocity, providing a relatively clear upper bound on this parameter. If $\Delta G_\tau^\dagger \lesssim 3$ $k_B T$ then translocation becomes very rapid and the same distribution of velocities is obtained in simulations, irrespective of the exact value of $\Delta G_\tau^\dagger$. In this case catalysis becomes strongly rate-limiting, and it would be appropriate to apply a partial equilibrium approximation to the translocation step. This provides an effective lower bound for parameter $\Delta G_\tau^\dagger$. Therefore we centered our prior distribution for $\Delta G_\tau^\dagger$ in this interval (a normal distribution with a mean of 5.5 and a standard deviation of 0.97, so that the central 99% interval is (3, 8)). We performed the same analysis with different values of K_D, as well as varying $\Delta G_{\tau 1}$, and arrived at the same interval for $\Delta G_\tau^\dagger$ (results not shown).

T7 pol: the same analysis was performed, however with $\Delta G_\tau^\dagger$ at its prior mean of -3.3 $k_B T$ (S2 Fig).

Prior for k_{bind}, which governs the rate of NTP binding

To select a prior for k_{bind} we performed similar simulations, but instead used Model 2 - the simplest kinetic binding model. k_{bind} and k_{cat} were sampled uniformly from relevant ranges, K_D was set to 100 μM and [NTP] = 1000 μM. (S2 Fig).

Depending on the exact value of k_{cat}, if $k_{bind} \lesssim 0.1$ μM^{-1} s^{-1}, then it is impossible to achieve a realistic velocity, providing a relatively clear lower bound on this parameter. If $k_{bind} \gtrsim 5$ μM^{-1} s^{-1} then binding becomes very rapid and the same distribution of velocities is obtained in simulations, irrespective of the exact value of k_{bind}. Again this is because catalysis becomes strongly rate limiting in this region, and it would be appropriate to apply a partial equilibrium approximation to the binding step. Hence we centered our (lognormal) prior around the interval (0.01, 5) - the conservatively selected bounds reflecting that the experimental data has been collected at differing NTP concentrations, altering the rate. Performing the same analysis with different parameters gave us a similar prior (results not shown).

Prior distribution related to rate of NTP release

Most previous estimates of K_D have been obtained using models which treated binding as an equilibrium process. This rapid-binding assumption restrains the values which K_D may take. Resulting estimates for K_D are typically in the order of 10^1 - 10^2 μM. However for a model in which binding is assumed to be slow estimates of $\frac{k_{rel}}{k_{bind}}$ can be lower, without compromising the fit to the data. This is apparent in simulations carried out using Model 2. S2 FigD shows that there is non-identifiability between $\frac{k_{rel}}{k_{bind}}$ and k_{bind}, with vastly different values of these two parameters producing the same output velocity. This has been demonstrated by Mejia et al. 2015 [41] who estimated $\frac{k_{rel}}{k_{bind}}$ to be 0.6 μM when NTP binding was rate-limiting.

The prior distribution for $\frac{k_{rel}}{k_{bind}}$ must permit all of these possibilities to be tested fairly during Bayesian inference. We therefore centered our lognormal prior for $\frac{k_{rel}}{k_{bind}}$ around a very broad range, with a central 95% interval of (0.2, 200). It is noted that selecting a prior distribution which does not discriminate between the kinetic and equilibrium binding models *a priori* may not be plausible and the inferences that can be made about this part of the model are correspondingly weak.

S1 Fig. Translocation transition state approximations. Three methods for estimating the Gibbs energy of basepairing in the transition state between translocation states $S(l, 0)$ and $S(l, 1)$. The effects of model parameters are not displayed here; these basepairing Gibbs energy terms would be added onto the red lines in Fig. 3.

S2 Fig. Simulations of the elongation pathway. Each point is a single simulation of the full *rpoB* gene (4029 nt). For (A-C), Parameters on the x- and z-axis are sampled uniformly at random from the displayed range at the beginning of each trial. The y-axis of each plot (mean elongation velocity) is then measured from the respective simulation. [NTP] and F held constant at 1000 μM and 0 pN respectively. (A) and (B): Relationship between $\Delta G_{\tau}^{\dagger}$ and k_{cat} for the melting model with binding at equilibrium (Model 8). $\Delta G_{\tau 1}$ set to its prior mean (0 for RNAP and pol II, and -3.3 for T7 pol). (C) Relationship between k_{bind} and k_{cat} for the kinetic binding model with translocation at equilibrium (Model 2). (D) Relationship between K_D and k_{bind} with translocation held at equilibrium (Model 2). K_D and k_{bind} sampled uniformly from specified range and velocity is measured. Samples with simulated velocities outside of the range 1-2 bp/s were discarded. [NTP] = 10 μM and k_{cat} = 100 s^{-1}.

S3 Fig. Posterior distributions of simulated velocities. Black open circles represent empirical mean pause-free velocities reported in the original publication [6, 27, 28]. Each blue/purple dot is a single sample simulated from the posterior distribution of parameters/models for the respective polymerase. 30 samples were taken from each of the three posterior distributions. For RNAP, $[\text{NTP}]_{eq}$ is defined as [ATP] = 5 μM, [CTP] = 2.5 μM, [GTP] = 10 μM, and [UTP] = 10 μM. Posterior distributions are for (A) RNAP, (B) pol II, and (C) T7 pol.

Acknowledgments

We wish to acknowledge the contribution of NeSI high-performance computing facilities to the results of this research. NZ's national facilities are provided by the NZ eScience Infrastructure and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure programme. URL https://www.nesi.org.nz.

References

1. Herbert KM, Greenleaf WJ, Block SM. Single-molecule studies of RNA polymerase: motoring along. Annu Rev Biochem. 2008;77:149–176.

2. Dangkulwanich M, Ishibashi T, Bintu L, Bustamante C. Molecular mechanisms of transcription through single-molecule experiments. Chemical reviews. 2014;114(6):3203–3223.

3. Sweetser D, Nonet M, Young RA. Prokaryotic and eukaryotic RNA polymerases have homologous core subunits. Proceedings of the National Academy of Sciences. 1987;84(5):1192–1196.

4. Sosunov V, Sosunova E, Mustaev A, Bass I, Nikiforov V, Goldfarb A. Unified two-metal mechanism of RNA synthesis and degradation by RNA polymerase. The EMBO journal. 2003;22(9):2234–2244.

5. Sousa R, Chung YJ, Rose JP, Wang BC. Crystal structure of bacteriophage T7 RNA polymerase at 3.3 Å resolution. Nature. 1993;364(6438):593.

6. Abbondanzieri EA, Greenleaf WJ, Shaevitz JW, Landick R, Block SM. Direct observation of base-pair stepping by RNA polymerase. Nature. 2005;438(7067):460–465.

7. Adelman K, La Porta A, Santangelo TJ, Lis JT, Roberts JW, Wang MD. Single molecule analysis of RNA polymerase elongation reveals uniform kinetic behavior. Proceedings of the National Academy of Sciences. 2002;99(21):13538–13543.

8. Bai L, Fulbright RM, Wang MD. Mechanochemical kinetics of transcription elongation. Physical review letters. 2007;98(6):068103.

9. Hodges C, Bintu L, Lubkowska L, Kashlev M, Bustamante C. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. Science. 2009;325(5940):626–628.

10. Galburt EA, Grill SW, Wiedmann A, Lubkowska L, Choy J, Nogales E, et al. Backtracking determines the force sensitivity of RNAP II in a factor-dependent manner. Nature. 2007;446(7137):820–823.

11. Skinner GM, Baumann CG, Quinn DM, Molloy JE, Hoggett JG. Promoter binding, initiation, and elongation by bacteriophage T7 RNA polymerase a single-molecule view of the transcription cycle. Journal of Biological Chemistry. 2004;279(5):3239–3244.

12. Dulin D, Vilfan ID, Berghuis BA, Hage S, Bamford DH, Poranen MM, et al. Elongation-competent pauses govern the fidelity of a viral RNA-dependent RNA polymerase. Cell reports. 2015;10(6):983–992.

13. Neuman KC, Abbondanzieri EA, Landick R, Gelles J, Block SM. Ubiquitous transcriptional pausing is independent of RNA polymerase backtracking. Cell. 2003;115(4):437–447.

14. Davenport RJ, Wuite GJ, Landick R, Bustamante C. Single-molecule study of transcriptional pausing and arrest by E. coli RNA polymerase. Science. 2000;287(5462):2497.

15. Tolić-Nørrelykke SF, Engh AM, Landick R, Gelles J. Diversity in the rates of transcript elongation by single RNA polymerase molecules. Journal of Biological Chemistry. 2004;279(5):3292–3299.

16. Abbondanzieri EA, Shaevitz JW, Block SM. Picocalorimetry of transcription by RNA polymerase. Biophysical journal. 2005;89(6):L61–L63.

17. Watson JD, Crick FH, et al. Molecular structure of nucleic acids. Nature. 1953;171(4356):737–738.

18. Maitra U, Nakata Y, Hurwitz J. The Role of Deoxyribonucleic Acid in Ribonucleic Acid Synthesis XIV. A Study of the Initiation of Ribonucleic Acid Synthesis. Journal of Biological Chemistry. 1967;242(21):4908–4918.

19. Erie DA, Yager TD, Von Hippel PH. The single-nucleotide addition cycle in transcription: a biophysical and biochemical perspective. Annual review of biophysics and biomolecular structure. 1992;21(1):379–415.

20. Rhodes G, Chamberlin MJ. Ribonucleic acid chain elongation by Escherichia coli ribonucleic acid polymerase I. Isolation of ternary complexes and the kinetics of elongation. Journal of Biological Chemistry. 1974;249(20):6675–6683.

21. Maoiléidigh DÓ, Tadigotla VR, Nudler E, Ruckenstein AE. A unified model of transcription elongation: what have we learned from single-molecule experiments? Biophysical journal. 2011;100(5):1157–1166.

22. Bai L, Shundrovsky A, Wang MD. Sequence-dependent kinetic model for transcription elongation by RNA polymerase. Journal of molecular biology. 2004;344(2):335–349.

23. Tadigotla VR, Maoiléidigh DÓ, Sengupta AM, Epshtein V, Ebright RH, Nudler E, et al. Thermodynamic and kinetic modeling of transcriptional pausing. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(12):4439–4444.

24. SantaLucia J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proceedings of the National Academy of Sciences. 1998;95(4):1460–1465.

25. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. The journal of physical chemistry. 1977;81(25):2340–2361.

26. Beaumont MA. Approximate Bayesian computation in evolution and ecology. Annual review of ecology, evolution, and systematics. 2010;41:379–406.

27. Schweikhard V, Meng C, Murakami K, Kaplan CD, Kornberg RD, Block SM. Transcription factors TFIIF and TFIIS promote transcript elongation by RNA polymerase II by synergistic and independent mechanisms. Proceedings of the National Academy of Sciences. 2014;111(18):6642–6647.

28. Thomen P, Lopez P, Bockelmann U, Guillerez J, Dreyfus M, Heslot F. T7 RNA polymerase studied by force measurements varying cofactor concentration. Biophysical journal. 2008;95(5):2423–2433.

29. Wang MD, Schnitzer MJ, Yin H, Landick R, Gelles J, Block SM. Force and velocity measured for single molecules of RNA polymerase. Science. 1998;282(5390):902–907.

30. Shaevitz JW, Abbondanzieri EA, Landick R, Block SM. Backtracking by single RNA polymerase molecules observed at near-base-pair resolution. Nature. 2003;426(6967):684–687.

31. Artsimovitch I, Landick R. Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals. Proceedings of the National Academy of Sciences. 2000;97(13):7090–7095.

32. Zhou Y, Navaroli DM, Enuameh MS, Martin CT. Dissociation of halted T7 RNA polymerase elongation complexes proceeds via a forward-translocation mechanism. Proceedings of the National Academy of Sciences. 2007;104(25):10352–10357.

33. Yin YW, Steitz TA. The structural mechanism of translocation and helicase activity in T7 RNA polymerase. Cell. 2004;116(3):393–404.

34. Yu J, Oster G. A small post-translocation energy bias aids nucleotide selection in T7 RNA polymerase transcription. Biophysical journal. 2012;102(3):532–541.

35. Dangkulwanich M, Ishibashi T, Liu S, Kireeva ML, Lubkowska L, Kashlev M, et al. Complete dissection of transcription elongation reveals slow translocation of RNA polymerase II in a linear ratchet mechanism. Elife. 2013;2:e00971.

36. Depken M, Galburt EA, Grill SW. The origin of short transcriptional pauses. Biophysical journal. 2009;96(6):2189–2193.

37. Lecca P. Stochastic chemical kinetics. Biophysical reviews. 2013;5(4):323–345.

38. Rambaut A, Drummond A. Tracer 1.6. University of Edinburgh, Edinburgh. UK. Technical report; 2013.

39. Larson MH, Zhou J, Kaplan CD, Palangat M, Kornberg RD, Landick R, et al. Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. Proceedings of the National Academy of Sciences. 2012;109(17):6555–6560.

40. Thomen P, Lopez PJ, Heslot F. Unravelling the mechanism of RNA-polymerase forward motion by using mechanical force. Physical Review Letters. 2005;94(12):128102.

41. Mejia YX, Nudler E, Bustamante C. Trigger loop folding determines transcription rate of Escherichia coli's RNA polymerase. Proceedings of the National Academy of Sciences. 2015;112(3):743–748.

42. Nedialkov YA, Nudler E, Burton ZF. RNA polymerase stalls in a post-translocated register and can hyper-translocate. Transcription. 2012;3(5):260–269.

43. KIreeva M, Trang C, Matevosyan G, Turek-Herman J, Chasov V, Lubkowska L, et al. RNA–DNA and DNA–DNA base-pairing at the upstream edge of the transcription bubble regulate translocation of RNA polymerase and transcription rate. Nucleic acids research. 2018;46(11):5764–5775.

44. Guajardo R, Lopez P, Dreyfus M, Sousa R. NTP concentration effects on initial transcription by T7 RNAP indicate that translocation occurs through passive sliding and reveal that divergent promoters have distinct NTP concentration requirements for productive initiation. Journal of molecular biology. 1998;281(5):777–792.

45. Arnold S, Siemann M, Scharnweber K, Werner M, Baumann S, Reuss M, et al. Kinetic modeling and simulation of in vitro transcription by phage T 7 RNA polymerase. Biotechnology and bioengineering. 2001;72(5):548–561.

46. Wong F, Dutta A, Chowdhury D, Gunawardena J. Structural conditions on complex networks for the Michaelis–Menten input–output response. Proceedings of the National Academy of Sciences. 2018;115(39):9738–9743.

47. Briggs GE, Haldane JBS. A note on the kinetics of enzyme action. Biochemical journal. 1925;19(2):338.

48. English BP, Min W, Van Oijen AM, Lee KT, Luo G, Sun H, et al. Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. Nature chemical biology. 2006;2(2):87–94.

49. Schnell S. Validity of the Michaelis–Menten equation–steady-state or reactant stationary assumption: that is the question. The FEBS journal. 2014;281(2):464–472.

50. Iost I, Guillerez J, Dreyfus M. Bacteriophage T7 RNA polymerase travels far ahead of ribosomes in vivo. Journal of bacteriology. 1992;174(2):619–622.

51. Vogel U, Jensen KF. The RNA chain elongation rate in Escherichia coli depends on the growth rate. Journal of bacteriology. 1994;176(10):2807–2813.

52. Ryals J, Little R, Bremer H. Temperature dependence of RNA synthesis parameters in Escherichia coli. Journal of bacteriology. 1982;151(2):879–887.

53. Tennyson CN, Klamut HJ, Worton RG. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. Nature genetics. 1995;9(2):184–190.

54. Darzacq X, Shav-Tal Y, De Turris V, Brody Y, Shenoy SM, Phair RD, et al. In vivo dynamics of RNA polymerase II transcription. Nature structural & molecular biology. 2007;14(9):796–806.

55. Mason PB, Struhl K. Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. Molecular cell. 2005;17(6):831–840.

56. Kainov DE, Lísal J, Bamford DH, Tuma R. Packaging motor from double-stranded RNA bacteriophage ϕ12 acts as an obligatory passive conduit during transcription. Nucleic acids research. 2004;32(12):3515–3521.

57. Makarova OV, Makarov EM, Sousa R, Dreyfus M. Transcribing of Escherichia coli genes with mutant T7 RNA polymerases: stability of lacZ mRNA inversely correlates with polymerase speed. Proceedings of the National Academy of Sciences. 1995;92(26):12250–12254.

58. Mejia YX, Mao H, Forde NR, Bustamante C. Thermal probing of E. coli RNA polymerase off-pathway mechanisms. Journal of molecular biology. 2008;382(3):628–637.

59. Burns CM, Richardson LV, Richardson JP. Combinatorial effects of NusA and NusG on transcription elongation and rho-dependent termination in Escherichia coli1. Journal of molecular biology. 1998;278(2):307–316.

60. Kingston R, Nierman W, Chamberlin M. A direct effect of guanosine tetraphosphate on pausing of Escherichia coli RNA polymerase during RNA chain elongation. Journal of Biological Chemistry. 1981;256(6):2787–2797.

61. Galburt EA, Grill SW, Bustamante C. Single molecule transcription elongation. Methods. 2009;48(4):323–332.

62. Kireeva ML, Nedialkov YA, Cremona GH, Purtov YA, Lubkowska L, Malagon F, et al. Transient reversal of RNA polymerase II active site closing controls fidelity of transcription elongation. Molecular cell. 2008;30(5):557–566.

63. Anand VS, Patel SS. Transient state kinetics of transcription elongation by T7 RNA polymerase. Journal of Biological Chemistry. 2006;281(47):35677–35685.

64. Usala SJ, Brownstein BH, Haselkorn R. Displacement of parental RNA strands during in vitro transcription by bacteriophage $\varphi6$ nucleocapsids. Cell. 1980;19(4):855–862.

65. Wu P, Nakano Si, Sugimoto N. Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation. The FEBS Journal. 2002;269(12):2821–2830.

66. Degen D, Feng Y, Zhang Y, Ebright KY, Ebright YW, Gigliotti M, et al. Transcription inhibition by the depsipeptide antibiotic salinamide A. Elife. 2014;3:e02451.

67. Voss NR, Gerstein M. 3V: cavity, channel and cleft volume calculator and extractor. Nucleic acids research. 2010;38(suppl_2):W555–W562.

68. Greive SJ, Von Hippel PH. Thinking quantitatively about transcriptional regulation. Nature Reviews Molecular Cell Biology. 2005;6(3):221–232.

69. Csilléry K, Blum MG, Gaggiotti OE, François O. Approximate Bayesian computation (ABC) in practice. Trends in ecology & evolution. 2010;25(7):410–418.

70. Brewer BJ, Foreman-Mackey D. DNest4: Diffusive Nested Sampling in C++ and Python. arXiv preprint arXiv:160603757. 2016;.

71. Ratmann O, Jørgensen O, Hinkley T, Stumpf M, Richardson S, Wiuf C. Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. PLoS Computational Biology. 2007;3(11):e230.

72. Richardson JP, Greenblatt J. Control of RNA chain elongation and termination. Escherichia coli and Salmonella: cellular and molecular biology. 1996;1:822–848.

73. Bonner G, Lafer EM, Sousa R. Characterization of a set of T7 RNA polymerase active site mutants. Journal of Biological Chemistry. 1994;269(40):25120–25128.