High Resolution Measurements of RNA Polymerase with Nanopore Tweezers

Ian C. Nova

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Jens H. Gundlach, Chair

Bertil Hill

Patrick Stayton

Paul A. Wiggins

Program Authorized to Offer Degree:
Department of Molecular Engineering and Sciences

University of Washington

Abstract

High Resolution Measurements of RNA Polymerase with Nanopore Tweezers

Ian C. Nova

Chair of the Supervisory Committee:
Professor Jens H. Gundlach
Physics

RNA polymerase (RNAP) is the molecular machine responsible for transcription, the process of making RNA from a DNA template. While the overall role of RNAP in this central cellular process is clear (RNAP moves along double stranded DNA, extending an RNA chain by 1 base with each step), questions remain about the timing and order of events within each reaction cycle [1]. With bulk biochemical assays that average over an ensemble of molecules, determining the dynamics of individual RNAP enzymes during transcription is difficult. For this reason, experiments that capture and observe single RNAP molecules have been particularly elucidating. Specifically, optical tweezers experiments, in which the position of single RNAP molecules along DNA are tracked over time, have helped develop and test various kinetic models for RNAP transcription [2]. By varying applied force and nucleotide triphosphate (NTP, the substrate for transcription) concentration, these experiments probe the energy landscape and reveal the coupling between chemical hydrolysis and mechanical motion [3]. However, in physiologically relevant conditions, single steps of RNAP occur too quickly and over too small a distance (1 base pair) to be detected by optical tweezers [4]. This necessitates averaging over multiple steps in the reaction cycle, prohibiting direct observation of individual enzyme states. In addition, although particular DNA sequences are known to interact specifically with RNAP, optical tweezers lacks the exact sequence registration, limiting methods of studying sequence - enzyme interactions at high resolution.

At the University of Washington, our group has helped develop a new single-molecule technology uniquely suited to study the details of RNAP dynamics. This method, Single-molecule Picometer Resolution Nanopore Tweezers (or SPRNT), is based off of the concept of nanopore DNA sequencing [5] and allows monitoring the translocation of single motor enzymes along DNA at unprecedented spatiotemporal resolution (40 picometers at millisecond time scales). In addition, each enzyme position measurement with SPRNT corresponds to a particular DNA sequence, allowing direct observation of enzyme interactions with DNA sequence. SPRNT has already been used to study motor proteins that walk along DNA: a DNA polymerase [6] [7] and a DNA helicase [8] [9].

In this thesis, I present my role in developing SPRNT to investigate RNAP during transcription. In section 0.1, I outline the relevant molecular systems and the techniques used to investigate them. I describe the development of SPRNT, starting with the fundamentals of nanopore sequencing, and compare this new technique to other methods. In section 0.2, I detail the methods developed to investigate *E. coli* RNAP with SPRNT, providing a guideline for future investigations of this enzyme. In section 0.3, I describe the initial results that were used to motivate and improve upon these methods. In section 0.4, I recount the relevant results of this investigation. I present the first measurements of single steps of *E. coli* RNAP during transcription at biologically relevant [NTPs] under an assisting force. I summarize how these measurements were used to track enzyme transitions between different states, developing a model for transcription. Using this model, I extrapolate the results obtained under various assisting forces to calculate the relevant rate constants at zero-force, providing a view into how RNAP behaves in its native environment. Next, I describe more SPRNT experiments that investigate RNAP pausing at particular sequences, including the first single-molecule detection of a ′half-translocated′ state during pausing. I detail a model for RNAP pausing resulting from this data. Finally, in section 0.5, I discuss the conclusions drawn from this work and the role SPRNT can play in future studies of transcription.

TABLE OF CONTENTS

0.1 Background

0.1.1 Watching one molecule at a time

Inside every cell, a host of 'molecular machines' coordinate and carry out complex tasks. As technologies to study and observe these systems improve, the details of the reactions are revealed, and our overall understanding of cellular mechanisms deepens. In particular, methods to determine the structure of biomolecules have had an large impact on our understanding of biological reactions at the molecular level. Techniques like x-ray crystallography can determine the atomic structure of individual proteins, and super resolution imaging technologies provide snapshots of many molecules interacting with impressive detail [10] [11]. However, these methods generally only provide static information. So, as our knowledge of molecular structure has deepened, our understanding of the dynamics (how these structures change over time) and functionality of these systems has typically lagged behind [1]. Bulk biochemical assays that can track the rate at which many enzymes simultaneously perform the same task can provide valuable insight into the "mean dynamics" of a biomolecular system, but obscure the underlying distributions that produce those averages.

For this reason, scientists have developed technologies to track single biomolecules over the course of a reaction. Single-molecule techniques have become the go-to methods for investigating dynamics in biochemistry [12]. Using technologies like Forster Resonance Energy Transfer (FRET), atomic force microscopy (AFM), magnetic tweezers, and optical tweezers, not only can average rates be calculated from many individual observations of single molecules, but the underlying distributions become available [13] [14] [15] [16]. From these experiments, metrics can be analyzed like the degree of heterogenity in a system (between individual molecules or in between reaction cycles for a given molecule), the probability distributions and time constants for individual reaction steps, and the probabilities of more rare or off-pathway events. In addition, some single-molecule methods (AFM, optical and

magnetic tweezers) can directly apply physical forces to the biomolecule under investigation, probing the energy landscape to investigate kinetic mechanisms.

In the early 1990s, single-molecule technologies were used to answer many open questions in the functioning of the biological proteins Actin and Myosin, responsible for rearrangement of the cytoskeleton [17] [18]. Using optical tweezers, ∼10 nM steps of these enzymes along micro tubules were directly observed, and the data helped uncover how both proteins couple the chemical hydrolysis of ATP to mechanical motion along the microtubule. With these studies, the idea of a molecular "machine" or "motor" was born; an enzyme that uses chemical energy for mechanical work. In the years since these experiments, single-molecule technologies have been used to study a variety of different molecular motors, and efforts to improve the throughput and resolution of these techniques have continued [1].

0.1.2 Single-molecule methods for DNA processing enzymes

Motor enzymes that interact with oligonucleotides (DNA and RNA, Figure 1) are an extremely important class of biomolecules. DNA polymerases, RNA polymerases, helicases, and ribosomes are paramount to every step in the central dogma of molecular biology (DNA to RNA to Protein) [19]. Generally, these enzymes catalyze the chemical hydrolysis of nucleotriphosphates (ATP, GTP, CTP, TTP, UTP) and couple the resulting energy release to drive mechanical work. The resulting functions are diverse and complex, and the mechanisms involved in these processes continue to be uncovered.

Single-molecule technologies have been employed to study basically every class of oligonucletoide processing enzyme, and the results have been illuminating. In particular, optical tweezers experiments have been useful in developing deeper kinetic understanding of these enzymes [1]. However, unlike motor proteins Actin and Myosin, which operate on the length scales of around 10 nm, olig0nucleotide processing enzymes move over much smaller distances throughout the course of a each reaction cycle. Specifically, the distance between base pairs on double stranded DNA (dsDNA) is ∼0.34 nm [20]. Enzymes that interact with every nucleotide along a molecule (polymerases), therefore, most likely move along DNA in steps

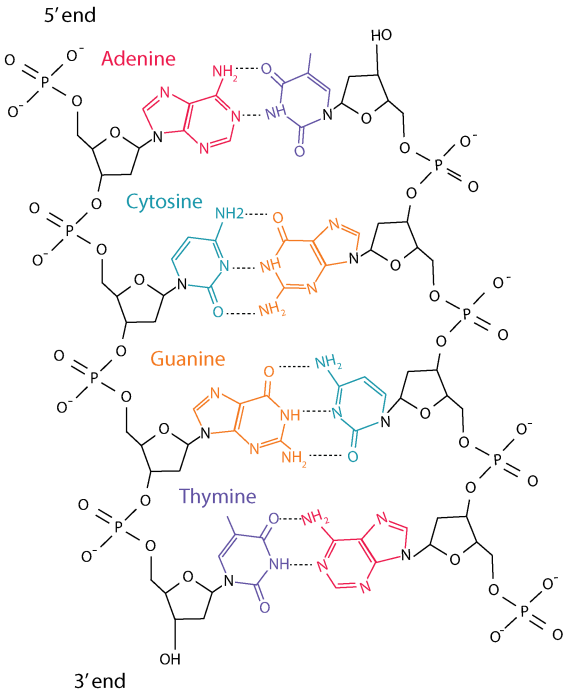


Figure 1: DNA (deoxyribonucleic acid) is a polymer containing a phosphate backbone and 4 unique nucleobases (Adenine (A), Thymine (T), Cytosine (C), Guanine (G)). Two strands of DNA run antiparrellel, forming hydrogen bonds with the complementary bases in the other strand (A:T, G:C)

close to the distance between bases in dsDNA. Although single-molecule technologies can track a polymerase along DNA, the spatiotemporal resolution of these techniques generally prohibits the identification of single enzyme steps [1].

In order to observe single steps of RNA polymerase, a dual optical trap was employed under low [NTP], slowing the rate of RNAP transcription to ∼1 step/second, and allowing sufficient time averaging to detect single nucleotide steps [4]. While this technique was useful for identifying the step size of the enzyme (1 bp), low NTP concentrations only allow observation of a single kinetic scenario in which the NTP binding rate limits the reaction velocity. Full investigation of enzyme kinetics requires observing reaction rates at a range of NTP concentrations. For higher [NTP], optical tweezers lacks the spatial resolution to identify single steps; so the data is averaged over multiple steps [4]. While much can be inferred from this type of analysis, absolute determination of the order of reaction steps and there relative rate constants will require observation of single reaction cycles (single steps of the enzyme) in varying conditions. These experiments will require a single-molecule method with improved spatiotemporal resolution.

In addition, studying sequence-specific behavior of motor enzymes can be difficult with conventional single-molecule methods. Each base within DNA and RNA (Adenine, thymine cytosine, guanine, and uracil) has a unique structure and energetic profile. Polymerases, helicases, and ribosomes interact with specific sequences in there olignucleotide substrates, coordinating their action and encoding for specific functions. Determining the details of how processing enzymes respond to sequence is vital for understanding each phase of the central dogma. Although optical tweezers can track a molecule along DNA or RNA, the exact DNA sequence inside the enzyme at any given time point is not determinable [1]. For this reason, it remains difficult study the sequence specific reactions of individual motor proteins in real time.

0.1.3 SPRNT: from DNA sequencing to enzymology

0.1.3.1 Nanopore DNA sequencing

Single-molecule Picometer Resolution Nanopore Tweezers (SPRNT) is a new technique based off of the principal of nanopore DNA sequencing [5]. In 1996, Kasianowicz et al. [21] first proposed the concept of using a nanopore to sequence DNA. Simply, nanopores are nanoscale diameter holes and exist in two classifications:

1) solid state nanopores: nanoscale holes manufactured into silicon nitride or graphene sheets

2) Biological nanopores: biological membrane porins

In a classic nanopore setup (Figure 2), a single pore in a membrane connects two wells (*cis* and *trans*) filled with a salt solution. When a voltage is applied across the membrane, positive and negative ions flow through the pore in opposite directions according to the electric field. This produces a measurable ion current characteristic to the type of nanopore. In nanopore sequencing, DNA is introduced into the *cis* chamber. Because DNA is polynegatively charged along its backbone, the molecule is drawn towards the positive *trans* chamber, and will flow through (translocate) through the nanopore. The translocation of DNA partially blocks the flow of ions, reducing the measured ion current. As the DNA molecule passes from *cis* to *trans*, the different nucleotides (A,T,G,C) within the molecule sequentially slide through the pore and block the current by different amounts [22]. Therefore, the recording of current during the translocation of a DNA molecule can be used to identify the sequence of that molecule. However, the rate at which DNA freely translocates through a nanopore is too fast to detect the changes in current produced by individual nucleotides (1nt/μs) [23]. In order to sequence DNA, the rate of DNA translocation must be precisely controlled. Achieving this slowed translocation was the bottleneck for full realization of nanopore DNA sequencing for nearly 15 years.

In the meantime, however, experiments with homopolymer DNA [22] and others with immobilized DNA inside a nanopore [24], [25], [26] helped further establish the proof of concept.

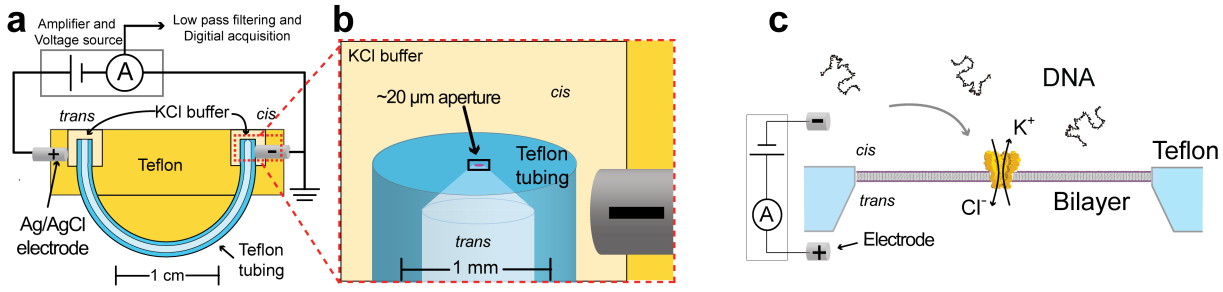


Figure 2: The typical nanopore sequencing experimental setup. A) Overview of the experimental apparatus. A teflon U-tube is embedded in a teflon puck. The U-tube is filled with salt solution and connects two wells (*cis* and *trans*) that contain Ag/AgCl reference electrodes. B) A ~20 μm aperture connects the end of the U-tube and the *cis* compartment. C) A lipid bilayer containing a single biological nanopore spans the aperture. Positive and negative ions flow through the nanopore according to the applied electric field. DNA in the *cis* chamber also moves according to the electric field towards *trans*, translocating through the nanopore during passage. This figure was adapted from [5] with permission.

By attaching sinlge-stranded DNA (ssDNA) to a large biomolecule like streptavadin that will not fit through a nanopore, DNA can be held statically inside the pore for an extended period of time. By comparing the measured ion currents for different DNA sequences, it was shown that homopolymers of each base produce unique currents, and that the currents changed with the substitutions of a single bases at certain positions [26]. In 2012, with the potential of nanopore sequencing fully established, methods to control translocation of DNA through a nanopore were first demonstrated [27] [6] [7]. In these schemes, a DNA motor enzyme, a protein that naturally translocates along DNA while performing a biological function, was loaded onto DNA and then introduced to a nanopore setup. The free end of a DNA molecule enters the pore and continues through until the motor enzyme (too large to fit through the nanopore) connects with the edge of the pore and translocation is halted. As the motor enzyme begins moving along DNA in single base steps, the DNA molecule is ratcheted through the nanopore with each step of the enzyme, allowing sufficient time for averaging the current at each DNA position (Figure 3 and Figure 4). Specifically, phi29 DNA polymerase was first

used as the motor enzyme in this scheme, pushing nanopore sequencing one step closer to becoming a viable commercial DNA sequencing option [6] [7].

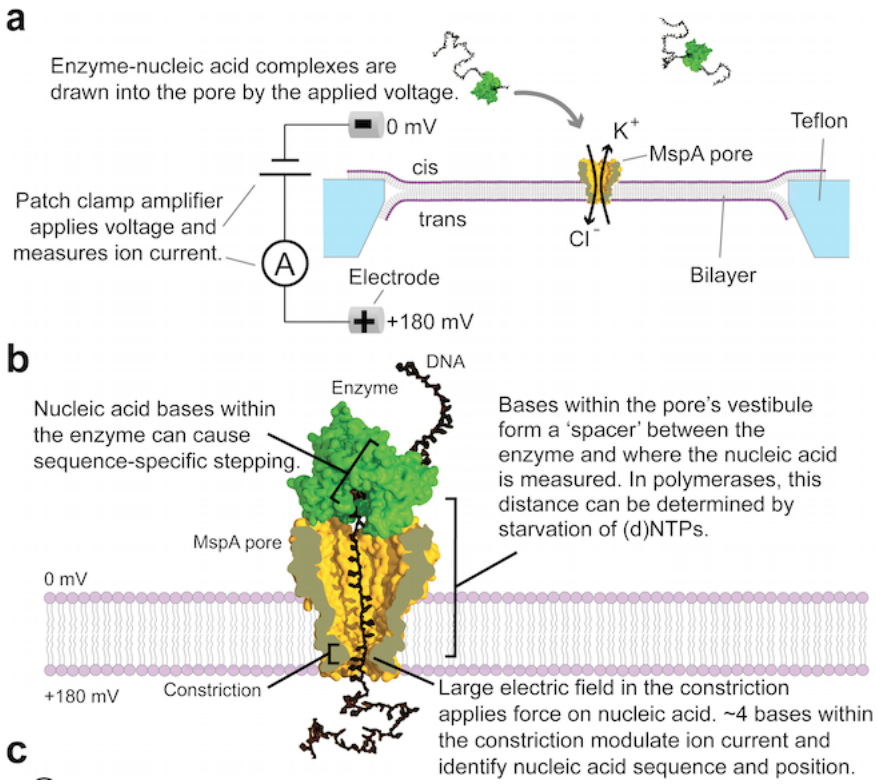


Figure 3: Experimental schematic for enzyme controlled DNA translocation through a biological nanopore. A) Motor enzyme loads onto DNA and the enzyme-nucleic acid complex is drawn into the pore by the applied voltage. B) The enzyme comes to rest upon the rim of the pore, and the stepping of the enzyme along the DNA substrate controls further translocation of the DNA through the nanopore. The DNA sequence within the constriction of the nanopore modulates the measured ion current through the nanopore. This figure was adapted from [5] with permission.

0.1.3.2 Nanopore sequencing with MspA

Biological nanopores (compared to solid state pores) have been most promising for nanopore DNA sequencing. Because proteins are atomically reproducible, the current patterns produced by the translocation of a particular DNA sequence do not change between biological nanopores of the same type. Solid state pores lack this reproducibility (they often vary in

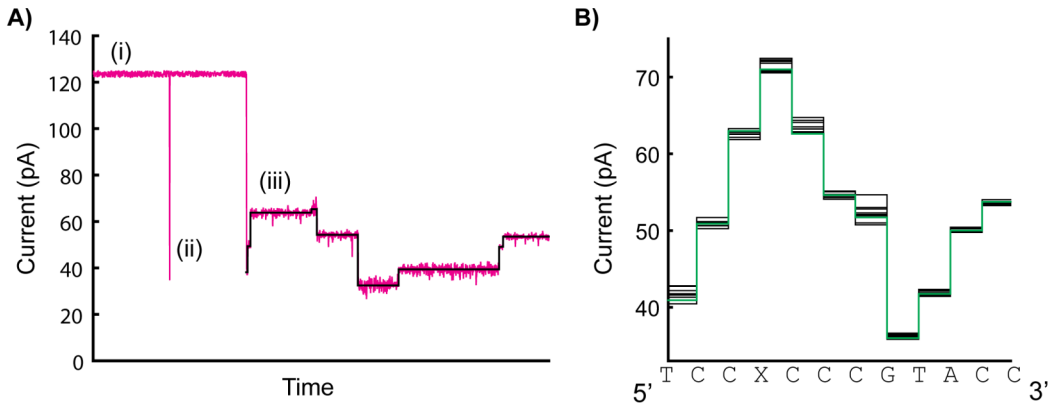


Figure 4: Example data trace from enzyme controlled DNA translocation through a biological nanopore (taken from [28]). A) The ion flow is recorded by measuring the current over time. (i) The unblocked pore produces a stable and reproducible current. (ii) Without the presence of a motor enzyme, DNA will translocate through the pore from *cis* to *trans* and reduce the measured current, but translocation occurs too quickly to resolve current changes produced by specific nucleotides in the DNA. (iii) With motor enzyme present, controlling DNA motion in the pore, discrete changes in current are observable with each step of the enzyme. A level finding algorithm identifies these discrete current states (levels) during controlled DNA translocation. B) The DNA sequence present in the pore during each translocation step determines the current pattern. After recording many controlled translocation events of the same DNA sequence and extracting the corresponding current levels, duration information resulting from stochastic enzymatic stepping behavior is removed. The individual current level patterns (black lines in (B)) are aligned and used to generate a consensus plot (green line in (B)) by taking the mean value of all the individual levels corresponding to each step. The DNA sequence producing the consensus current pattern is plotted. For each DNA translocation step, 4 nucleotides present in the constriction zone of MspA affect the current magnitude. X denotes an abasic residue.

size based on manufacturing), and will change shape over the course of an experiment. Biological pores alpha hemolysin (αHL) [3] [22] [24] and Mycobacterium Smegmatis Porin A (MspA) [23] [29] [26] have dimensions similar to the size of a DNA molecule, making them prime candidates for nanopore sequencing experiments. Here at the University of Washington, in the biophysics lab of Dr. Jens Gundlach, researchers pioneered early efforts to realize nanopore DNA sequencing with MspA, as the hourglass shape and the shorter (0.6 nm) and narrow (1.2 nm) constriction of the pore seemed ideal for these applications (Figure 5) [23].

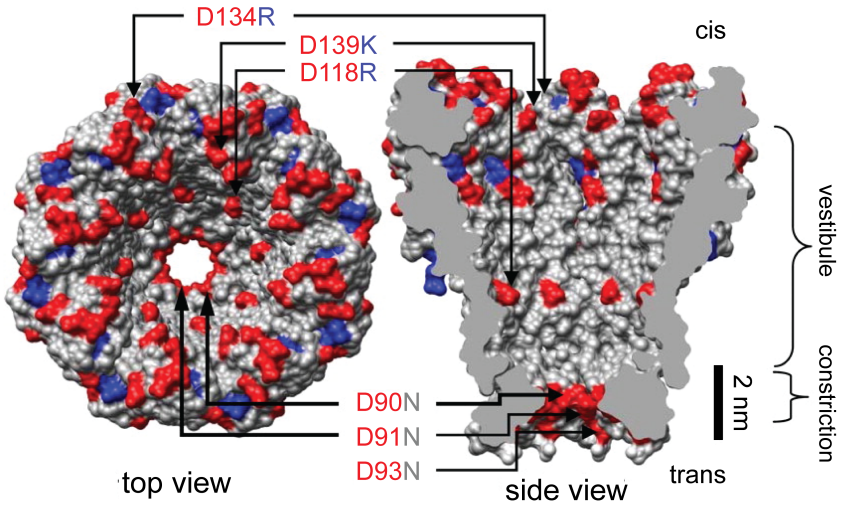


Figure 5: Stucture of engineered MspA nanopore. *Mycobacterium smegmatis* porin A (MspA) is an outer membrane porin of *Mycobatcerium smegmatis*. MspA was engineered for nanopore sequencing [23]. Negatively charged aspartic acid residues (D134,139,118,90,91,93) were replaced with either neutral (in the constriction) or positively charged (along the vestibule) residues. These mutations increase DNA capture into the pore and allow DNA translocation through the constriction. This figure was adapted from [5] with permission.

To begin, the nanopore lab at the University of Washington engineered MspA, replacing negatively charged residues within the constriction and along the rim of the pore to allow the translocation of DNA (Figure 5) [23]. Next, as noted above, experiments were performed to demonstrate the nucleotide sensitivity of MspA [26] and to control translocation of DNA through the pore using phi 29 DNAP [7]. Further studies determined that the phi29 DNAP-MspA method was sensitive enough to detect single methylation and hydroxymethylation sites on DNA [30] as well as unnatural DNA bases dNam and d5sics [31] These experiments revealed that, even though the size of the constriction of MspA is close to a single nucleotide, at any given time, $\sim$4 nucleotides contribute to the measured current due to thermally driven fluctuations in the positions of the nucleotides relative to the constriction during a single enzyme state. Therefore, each 4-base combination (or Quadromer) produces a specific ion current. Thus, with 4-base combinations and 4 DNA base types, there are 256

unique quadromers (4^4). The entire set of quadromers (quadromer map) was experimentally determined with MspA (Figure 6). This map can be used to predicted the measured current patterns for any DNA sequence, allowing accurate reference sequencing (where a nanopore DNA sequencing read is compared to a database of organismal genomes, and a match is found) [32].

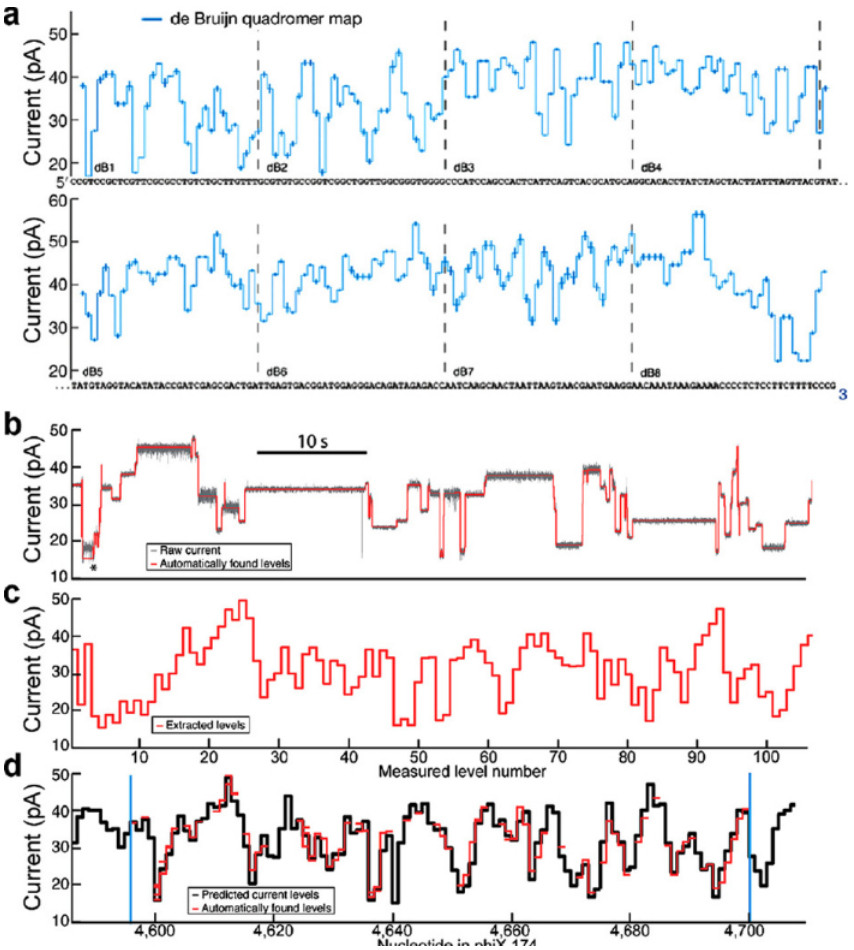


Figure 6: Reference sequencing with MspA and phi29 DNAP (adapted from [32] with permission.). A) The Quadromer Map (all 256 DNA sequence-current relationships) was determined using phi29 DNAP and MspA. B) Long nanopore reads of the phi X 174 bacteriophage genome were recorded and level found (C). D) The phi X 174 nanopore reads were aligned to the predicted current pattern of the genome determined using the quadromer map. Mismatches between the measured levels and the predicted levels are mostly due to 'skips' taken by the enzyme. T

Although the error rates for de novo sequencing (determining the DNA sequence without any reference) with nanopores are still high compared to other techniques, nanopore sequencing offers other advantages. Because nanopore DNA sequencing is inherently a single-molecule method, detection of modified DNA bases is possible [30]. These epigenetic modifications encode messages into DNA and modify gene expression. In addition, long molecules of DNA (10-100 kb) can be captured and sequenced using nanopores, while standard sequencing reads are only 250 bp. With short reads, genome reassembly (finding the location of the short DNA sequence along the entire genome) is often prohibitively difficult. Nanopore reads have been used to augment genome reassembly methods with shorter reads [32]. Current research to improve base calling accuracies and overall experimental throughput continue in our lab and elsewhere, including multiple commercial ventures [33] [34].

0.1.3.3 Single-molecule enzymology with nanopores

Endeavors to sequence DNA with nanopores led to a coincidental discovery: the same experimental setup could be used to investigate the dynamics of single enzymes that walk. When the rate of DNA translocation through a nanopore is controlled by an enzyme, the measured ion currents reflect the motion of the enzyme along DNA. By recording many reads of an enzyme navigating along a known DNA sequence, information is acquired about the rate of enzyme stepping at each DNA position as well as the stepping behaviour (backsteps, skips, etc..) and step size. [5]. In addition, the applied electric field driving DNA translocation applies a force to the motor enzyme. Varying the applied voltage will vary the applied force, allowing study of the kinetic response to a range of applied forces. Experiments using phi 29 DNAP to control DNA translocation through nanopores provided direct insight into the mechnism of action by phi 29 in polymerizing DNA [35] [36]. Other enzymes have been studied in this fashion. Specifically, experiments with hel 308 DNA helicase, an enzyme that unwinds dsDNA, have further demonstrated the utility of this analysis technique [8], identifying steps smaller than the distance between single base pairs and cal-

culating reaction rates specific to DNA sequences. These discoveries have demonstrated the utility of using nanopores as an enzymology tool, and have led to a general technique, named **Single-molecule Picometer Resolution Nanopore Tweezers (SPRNT)**. SPRNT can be applied to study many DNA and RNA motor enzymes in unprecedented detail.

0.1.3.4 From ion current to enzyme position

When DNA is pulled through MspA by a motor enzyme, the resulting ion current trace contains discrete current measurements, sampling the DNA at given intervals based on the step size of the enzyme (Figure 7a). For an enzyme that takes single-nucleotide steps (like phi 29 DNAP), the discrete current pattern represents sampling of the DNA at single nucleotide intervals. Hypothetically, if a DNA strand was pulled continuously through the pore, the measured current would be a smooth curve, connecting the discrete current measurements. These smooth curves for a given DNA sequence can be constructed by interpolating a spline to the single nucleotide ion current measurements for the same sequence (Figure 7b). These smooth DNA position vs current curves can be used as a ruler to measure the DNA position at any given time. For example, Hel 308 helicase takes 1/2 nucleotide distance steps along DNA. Aligning the discrete ion current patterns generated by hel 308 to the smooth position vs current trace for the same sequence (Figure 7c and d) assigns a DNA position for each discrete step in the hel 308 trace (Figure 7e). The resulting position vs time trace (Figure 7e) for hel 308 shows discrete steps at every half nucleotide position. This method can be applied to any new motor enzyme being studied with SPRNT.

0.1.3.5 Comparing SPRNT to other methods

SPRNT is extremely sensitive to small changes in DNA position within the constriction of MspA. In fact, position uncertainty as small as 0.06 nt can be resolved [8], corresponding to a distance uncertainty of $\sim$40 picometers for milisecond long steps. The noise in any SPRNT current measurement is Gaussian-distributed. For this type of measurement, the spatiotemporal resolution follows a -1/2-power law relation between observable step size and

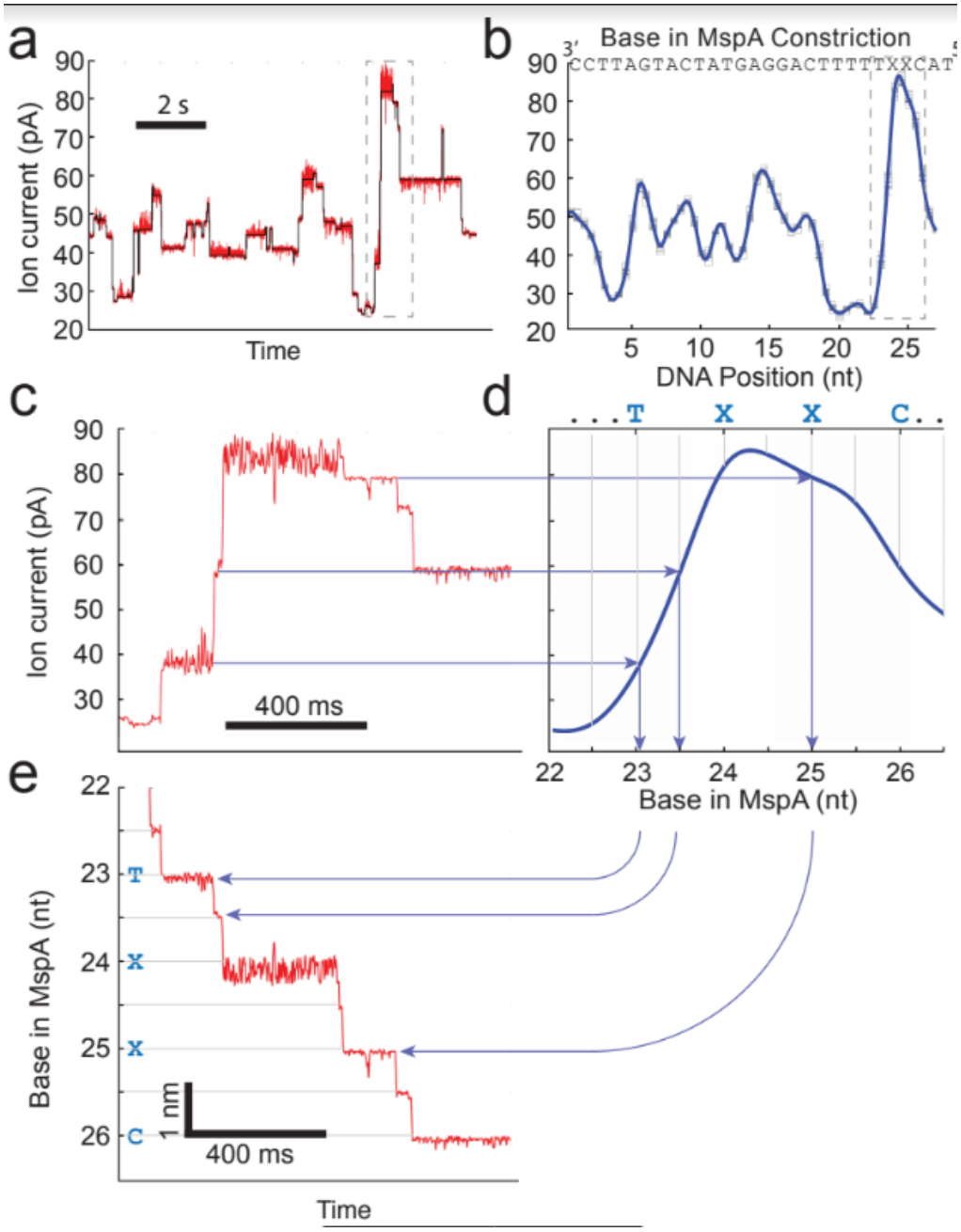


Figure 7: SPRNT: Ion current vs time is converted into DNA position vs time. A) Enzyme controlled DNA translocation is used to generate an ion current current vs time trace for a specific DNA sequence. B) A smooth ion current vs DNA position trace is generated by interpolating between discrete current vs DNA position values. C) and D) An ion current vs time trace is aligned to the smooth ion current vs DNA position trace. Every data point in ion current space is assigned a DNA position value. D) The resulting DNA position vs time trace is a map of a single motor enzyme's location along DNA in real time and can be used to analyze the dynamics of translocation of the motor enzyme. In this instance, hel308 DNA translocase was used, so the position vs time trace contains discrete steps at half nucleotide intervals (the step size of hel308 translocase)

Technique	sSNR	Force(pN)	Distance range (nm)	torque?	Massively Parallelizable ?
SPRNT	2360	15-60	0.04 - 10^5	No	Yes
OT	41.6	0.1 - 100	0.1 - 10^5	Yes	No
MT	24.3	0.001 - 10000	0.5 - 10^5	Yes	Yes
TIR-FRET	41.6	–	2 - 10	No	Yes

Table 1: Adopted from [5]. A comparison of the properties of several single-molecule techniques. SPRNT has a superior signal-to-noise ratio (SNR) and can measure enzyme movement over the smallest distance ranges. However, the applied force range and lack of ability to apply torque to the system under investigation may limit SPRNT in specific applications.

observable duration [15]. Figure 8 compares the spatiotemporal resolution of SPRNT to other single-molecule methods. It is important to note that the step sizes and durations of many important biological motor enzymes fall below the limits of detection of all technologies except SPRNT. Other factors are also to be considered when assessing single-molecule techniques (Table 1). One caveat of SPRNT is that the applied force is uncalibrated; i. e. the conversion between applied voltage (mV) and force (pN) is based on estimates and has not been directly calculated [8]. For SPRNT to reach its full potential, direct measurements of the applied forces are needed.

0.1.3.6 Determining the enzyme location in SPRNT

As noted previously, SPRNT holds promise in studying the sequence-specific behavior of motor enzymes. SPRNT measures the position of DNA in the constriction of MspA with high precision, but the exact position of the enzyme relative to the DNA in the pore must be experimentally determined. To phrase this differently, the distance between the sequence in the active site of the enzyme and the sequence in the constriction of the pore must be calculated (Figure 3b). This 'spacer correction' will be unique to each motor enzyme studied with SPRNT. The spacer correction should be relatively constant for a given enzyme, and can be simply added to the DNA position measurement throughout the course of a SPRNT

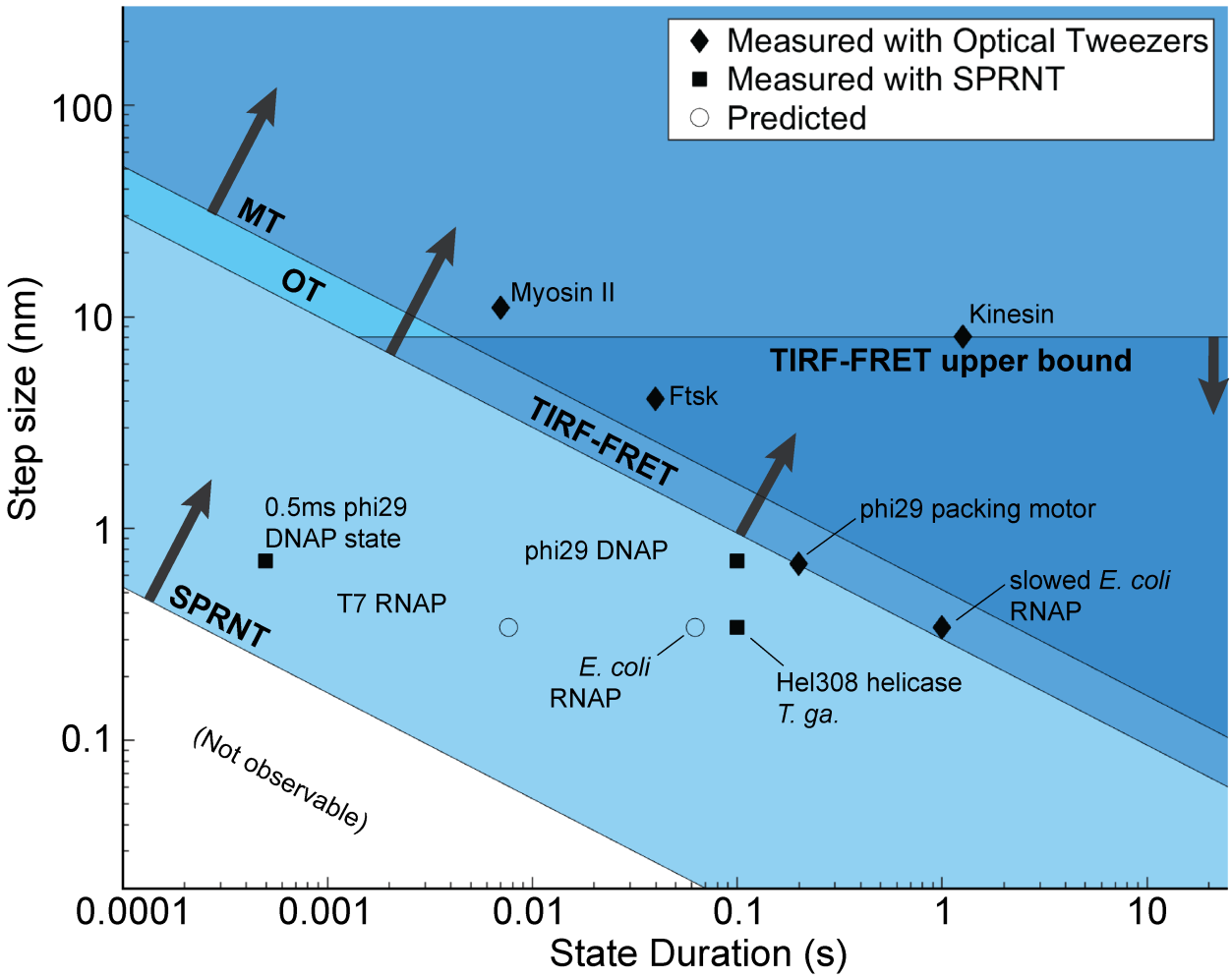


Figure 8: Comparsion of Single-molecule Methods. The spatiotemporal resolution of single-molecule technologies are compared (Magnetic Tweezers (MT) , Optical tweezers (OT), Total Internal Reflection Fluorescence Forster resonance energy transfer (TIRF-FRET), and Single-moleculee Picometer Resolution Nanopore Tweezers (SPRNT)). Spatiotemporal resolution follows a -1/2-power law relation between observable step size and observable duration. The step size and step durations for important motor enzymes are plotted. Enzymes above and to the right of the spatiotemporal cutoff for each technique can be observed. This figure was adapted from [5] with permission.

experiment to determine the enzyme location; although, it is important to note that possible movements or restructuring of the enzyme subunits at any given DNA position could effect the DNA position within the pore (changing the spacer correction).

Calculating the spacer correction is possible for enzymes with various unique reactant species. For example, phi29 DNAP uses four unique dNTPs to synthesize DNA (dATP, dTTP, dCTP, dGTP). By decreasing one of the four reactant concentrations (low [dATP], high [dTTP, dCTP, dGTP]), steps requiring the incorporation of an dATP will increase in duration. By comparing the relative duration patterns over many nanopore reads for a series of steps, the pattern of A's in the DNA can be effectively sequenced. This method can be repeated for all four dNTPs. Comparing the pattern of durations to the DNA sequence within the nanopore with each step (determined by ion current magnitude) can determine the distance between the enzyme active site and the constriction [35]. Knowing the exact sequence within the active site of the enzyme during each SPRNT measurement allows investigation of sequence-specific behavior unlike any other single-molecule method.

0.1.4 RNA Polymerase: the transcriber

RNA Polymerase is a complex molecular motor protein that uses nucleotide triphosphates (NTPs) to sequentially build a single stranded RNA molecule complementary to a DNA template. Transcription is generally separated into three phases:

(1) initiation: RNAP recognizes a specific DNA sequence (promoter), loads onto the DNA, opens up a bubble in the double-stranded DNA (dsDNA) (transcription bubble), and begins building an RNA transcript using individual nucleotide triphosphates (NTPs).

(2) elongation: after escaping the promoter region, RNAP translocates along the dsDNA, moving 3' to 5' along the template strand and elongating the RNA transcript by one base with each step.

(3) termination: Upon encountering a specific protein factor (rho-dependent) or a specific DNA sequence (intrinsic), RNAP decouples from the DNA and transcription ends.

Questions remain about each phase of transcription. In the work proposed here, I will

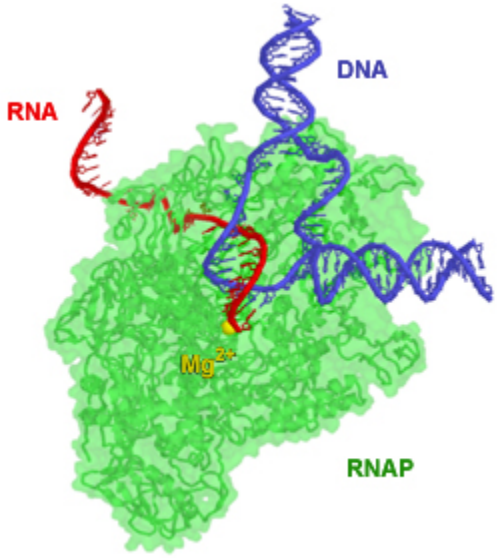


Figure 9: Schematic of RNAP during transcription. T7 RNAP [37] is a single subunit RNAP that transcribes RNA from a DNA template. A bubble in dsDNA (transcription bubble) is opened inside the core of the enzyme. The enzyme translocates along dsDNA, shifting the transcription bubble downstream and catalyzing the addition of RNA bases to the 3' end RNA transcript. Image acquired from wikipedia (public domain)

investigate mechanisms involved in transcription elongation and intrinsic termination. With the ability to track RNA Polymerase at high resolution and with exact sequence registration, SPRNT is ideally suited for this study.

0.1.4.1 E. coli RNA Polymerase: structure and mechanism

Molecular interactions in *E. coli* are investigated and treated as model systems: simpler representatives of systems conserved in higher orders of life. For RNAP, the structure and general mechanisms are highly conserved throughout both prokaryotes and eukaryotes [38]. *E. coli* RNAP is a 500 kDa enzyme composed of subunits $\alpha_2\beta'\beta\omega\sigma$ and can exist in two forms. The holoenzyme refers to the whole RNAP complex in the presence of the sigma factor subunit; known for its role in transcription initiation. The RNAP core enzyme refers to the complex in the absence of sigma factor and can processively elongate RNA. Crystal structures of RNAP core from *Thermus aquaticus* (structurally similar to *E. coli* RNAP

core) reveal the organization of the transcription elongation complex (TEC) [10]. The TEC contains a transcription bubble, where 12 DNA:DNA base pairs between the template and nontemplate strand are melted, and the template DNA strand instead hybridizes with the growing RNA chain through 9-10 DNA:RNA contacts (see Figure 10). An NTP entry channel allows NTPs in solution access to the NTP incorporation site deep within the protein. In each reaction cycle, the RNAP catalyzes the addition of one nucleotide to the 3' end of the RNA, complementary to the sequence along the template strand.

Structural and biochemical analysis has demonstrated that, during transcription elongation, the TEC alternates between two states (pre-translocated and post-translocated, Figure 10 [39]. In the pre-translcoated state (TEC_{pre}), the NTP incorporation site is filled by the 3' end of the RNA transcript, and 10 DNA:RNA bases pair between the RNA and the DNA template strand. In transitioning to the post-translocated state (TEC_{post}), the transcription bubble shifts downstream by one base, breaking one DNA:DNA base pair on the downstream edge, breaking one RNA:DNA base pare on the upstream edge, and adding one DNA:DNA base pair on the upstream edge. This leaves the NTP incorporation site empty, and the length of the RNA:DNA duplex at 9 bp, one less than during TEC_{pre}. In order to return to TEC_{pre}, an NTP binds in the active site and is incorporated onto the 3' end of the RNA [12]. It is important to note that while the position of RNAP moves forward by 1 bp with respect to the DNA when transitioning from TEC_{pre} to TEC_{post}, the next transition from TEC_{post} to TEC_{pre} does not include a physical translocation of the enzyme. Therefore, for every physical position of the enzyme during transcription, the TEC may exist in either pre or post states. To avoid confusion in writing chemical reaction networks for transcription elongation, the length of the RNA transcript (n,n+1,n+2,...) for each TEC state is labeled. The simplified reaction network below depicts 1.5 reaction cycles. The transitions that correspond to enzyme translocation are labeled *trans*:

$$\ldots\ TEC_{pre,n} \overset{trans}{\rightleftharpoons} TEC_{post,n} \rightleftharpoons TEC_{pre,n+1} \overset{trans}{\rightleftharpoons} TEC_{post,n+1}\ \ldots$$

Scientists in the field of RNAP research generally agree upon this overall mechanism and the structures of these translocation states. However, disagreements and alternative

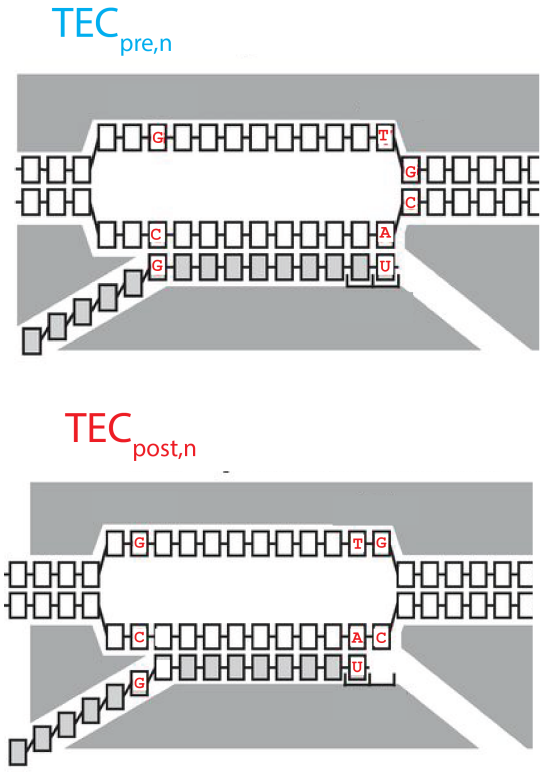


Figure 10: Two translocation states of RNAP. The transcription elongation complex (TEC) consists of a transcription bubble in dsDNA (white) and an RNA transcript (grey) hybrdized to the DNA template strand. The grey block surrounding the transcription bubble represents the RNAP. An NTP entry channel allows the passage of NTPs into the enzyme (bottom right). The RNA transcript exits through the RNA exit channel (bottom left). During transcription elongation, the TEC alternates between two translocation states: the pretranslocated state (TEC_{pre} , blue) and the posttranslocated (TEC_{post} , red).

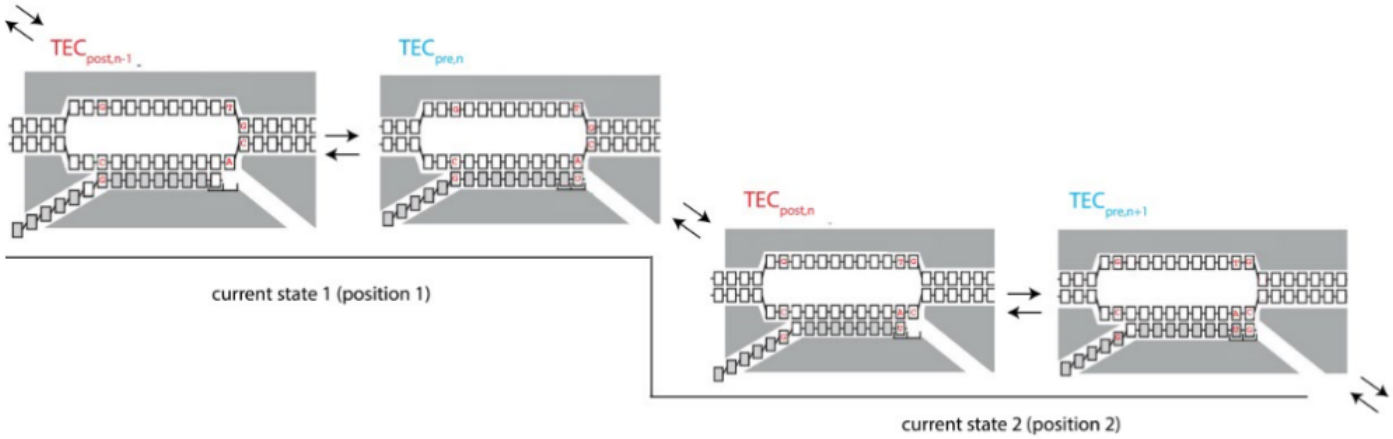


Figure 11: Translocation states of RNAP during SPRNT. The position of RNAP moves forward by 1 bp with respect to the DNA when transitioning from TEC_{pre} to TEC_{post}. The next transition from TEC_{post} to TEC_{pre} does not include a physical translocation of the enzyme. Therefore, each translocation state observed in a single-molecule assay (ion current states when using SPRNT) corresponds to both a $TEC_{post,n}$ to $TEC_{pre,n+1}$, where n denotes the length of the RNA transcript.

hypothesis arise when mapping out the order of reaction steps within a given reaction cycle and when calculating the relative magnitudes of the reaction rates for each step within the reaction. Additionally, experimental evidence has suggested the existence of both forward-tracked and backtracked TECs, in which the enzyme moves "off-pathway" and physically translocates either forward or backward without synthesizing RNA [40] [41]. RNA elongation only continues when the enzyme returns from the off-pathway state, and the 3' end of the RNA is returned to the active site of the enzyme [12]. While important to understanding the whole picture, we will initially ignore this off-pathway behavior for the discussion of kinetic models of transcription elongation.

0.1.4.2 RNA Polymerase: insights from optical tweezers

The first single-molecule experiments on RNAP were performed in 1995, when Yin *et al.* used optical tweezers to measure the force-velocity relationship of single RNAP complexes during elongation [2]. From these initial studies, stuctural and biochemical data, and subsequent

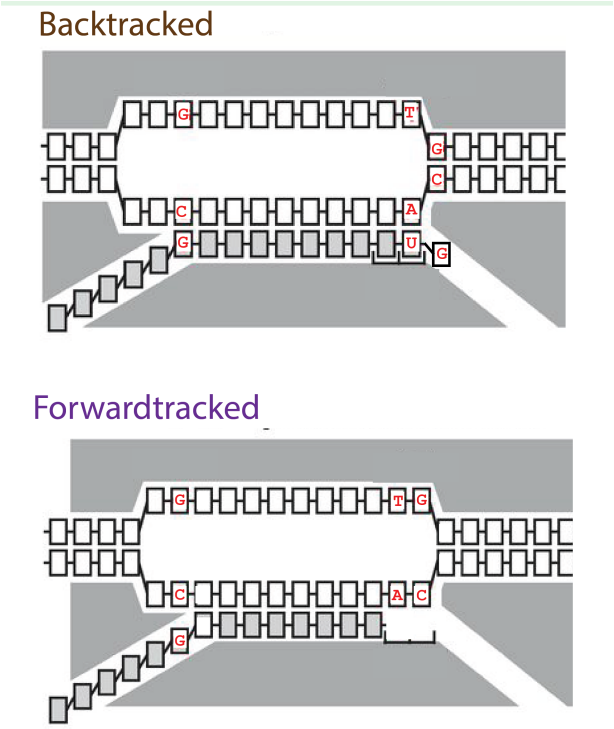


Figure 12: Off-pathway translocation states of RNAP. The TEC can exist in off-pathway conformations. In the backtracked state, the 3' end of the RNA transcripts extends into the NTP entry channel. In the forwardtracked state, the transcription bubble shifts forward before the incorporation of the next NTP, removing an extra base pair in the RNA:DNA hybrid

single-molecule efforts by the groups of Carlos Bustamante, Steven Block, and Michelle Wang over the next decade, an overall picture of transcription elongation was built. Some initial observations were consistent. RNAP is a processive enzyme that moves in a discontinuous fashion, alternating between a relatively constant velocity and a second, slower velocity around 0 bp/s, refered to as RNAP pausing [42] [43]. Instantaneous velocity distributions for individual RNAP molecules showed this bimodal distribution, fit well by the sum of two Gaussians [44]. While the non-pausing velocity for a given molecule was consistent along the DNA template, the velocities between different RNAP molecules in the same conditions differed, demonstrating a degree of variation between the inherent rates of TEcs (dynamic disorder). By applying opposing forces to the transcription complex, the stall force was determined in the range of 15-25 pN [42] [43] [42] [44]. Interestingly, some of these early optical tweezers studies suggested that transcription velocity was independent of force below the stall force [44]. However, more recent studies have refuted this, showing a velocity-force dependence over a wide range of forces and [NTPs] [45] [46] [4]. Applying a force will only affect the rates in the reaction that correspond to translocation. These force-velocity relationships have been used to develop kinetic models for transcription elongation.

0.1.4.3 Kinetic models of RNAP transcription elongation

Single-molecule experiments allow pauses in elongation to be detected and separated from "normal" elongation. Although pausing is an important part of the elongation mechanism, research has suggested it is an "off-pathway" state and occurs relatively infrequently (1 pause per 100 bases) [44]. For this reason, removing pausing is justified for determining the kinetic mechanisms of "on-pathway" transcription elongation.

Kinetic modeling of transcription elongation can be divided into two categories: (1) how chemical catalysis is coupled to mechanical motion (the translocation between $\mathrm{TEC_{pre}}$ and $\mathrm{TEC_{post}}$) and (2) how DNA sequence affects the kinetics

(1) Mechano-chemical coupling: models for molecular motors are broadly described based on how tightly coupled the chemical catalysis is to the mechanical motion. In one class of

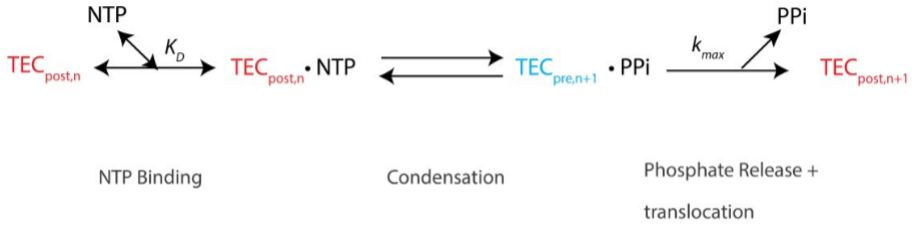


Figure 13: Power Stroke model of transcription elongation (no off-pathway behavior is included)

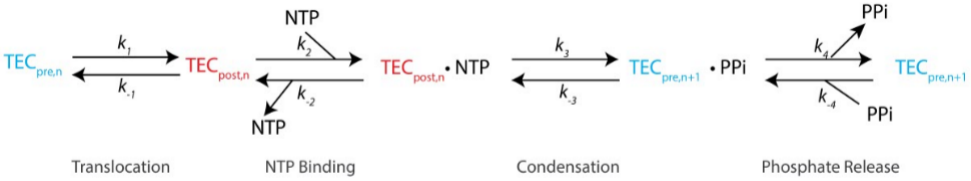


Figure 14: Brownian ratchet model of transcription elongation (no off-pathway behavior is included)

kinetic models, known as the "Power Stroke", physical translocation of the enzyme is driven by a chemical step in the pathway (tight coupling of catalysis and translocation). Experimental evidence supporting this model for transcription elongation is minimal. However, structural studies of T7 RNAP did suggest that forward translocation is driven by the release of pyrophosphate (PPi), fitting with a Power-Stroke mechanism (Figure 13) [47]. In the second class of models, translocation is thermally driven. The polymerase oscillates between $_{\text{pre}}$ and TEC_{post} until an incoming NTP rectifies TEC_{post}. This is referred to as a "Brownian ratchet" mechanism (Figure 14) [48]. Combined structural and biochemical analyses over the course of a decade has supported the Brownian ratchet mechanism. [49] [50] [51] [52].

Single-molecule evidence has backed up the Brownian ratchet mechanism [4] [46] [45]. In particular, the strong dependence of translocation velocity on force at low [NTPs] shown by Abbondanzieri et. al deviates from the Power Stroke model [4]. With low [NTPs], the NTP binding becomes rate limiting. Because force only affects reaction steps in which translocation occurs, and translocation is irreversible in a power stroke model, the NTP

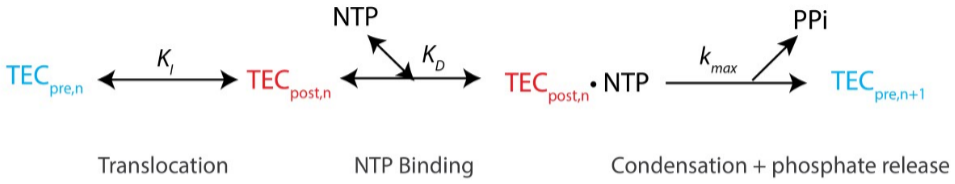


Figure 15: Brownian ratchet model of transcription elongation (assuming rapid equilibrium and low [PP_i])

binding step should be unaffected by force. Thus, when [NTP] binding is the rate limiting step at low [NTP], a Power Stroke model predicts velocity should be force-independent (see Figure 13). Overall, between single-molecule, biochmeical, and structural support, a Brownian ratchet mechanism has become generally accepted.

However, details of the Brownian ratchet mechanism pathway stemming from single-molecule studies have continue to be debated. In attempts to fit this model to experimental results, it has often been assumed that the translocation step and NTP binding steps occur in rapid equilibrium relative to the chemical steps in the cycle (condensation + phosphate release) [48] [4] [46] [50] [53]. This leads to the simplified reaction pathway shown in Figure 15, where equilibrium constants are used for the first two steps and the chemical steps are combined into one rate-limiting step (k_{max}). The last step is treated as irreversible, as [PP_i] is low in relevant conditions. While this model was used to accurately fit and then predict optical tweezers data for *E. coli* RNAP [46], other data acquired with the same enzyme over a greater range of forces could not be accurately described [4] [45]. Instead, Abbondanzieri et. al proposed a "branched" version of a Brownian ratchet (Figure 16). In this model, an NTP can bind in either translocation position TEC_{pre} or TEC_{post}. Although this model accurately describes single-molecule data [4] [45], it necessitates the existence of a secondary NTP binding site, which may exist, but lacks experimental evidence [54].

Alternatively, most of the single-molecule data can be described by challenging the assumption of rapid equilibrium between the translocation states [54] [55]. It is important to note that while widely assumed, rapid equilibrium has never been determined experimen-

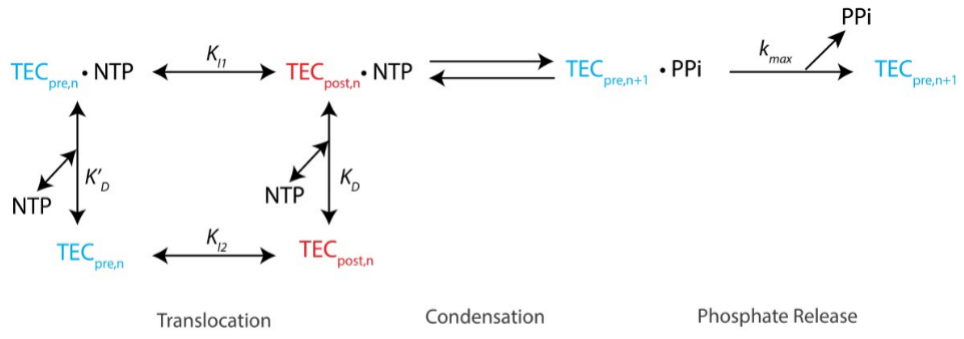


Figure 16: Branched Brownian ratchet model of transcription elongation (two NTP binding sites)

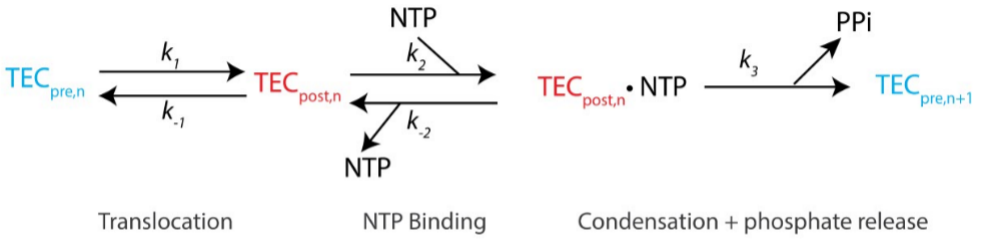


Figure 17: Brownian ratchet model of transcription elongation (non-rapid equilibrium)

tally. Instead, Dangkulwanich et. al started with the reaction model in Figure 17, and derived and fit each of the rate constants to their single-molecule data. They found that the forward translocation (k_1) rate is the same order of magnitude as the catalysis rate (k_1). This model can be used to explain the force-velocity relationships leading to the branched Brownian ratchet mechanism (Figure 15) [4] [45].

(2) Sequence specific kinetics

The reaction schemes in the previous section treated each reaction cycle as an identical process. However, every rate constant will be affected by the TECs position with respect to DNA sequence. Sequence-dependent kinetic models have been formulated by Tadigotla et. al using structural models [53] and Bai et.al from optical tweezers experiments [46] [50]. In the Brownian ratchet mechanism (assuming rapid equilibrium Figure 15), the rates of translocation between TEC_{pre} and TEC_{post} are governed by the difference in free energy

between the two states (Equation 1).

$$K_{\mathrm{i}}(n) = e^{(\Delta G_{post,n} - \Delta G_{pre,n})/(k_b T)} \tag{1}$$

The value of this free energy at each position (n) varies with the underlying DNA sequence inside the enzyme at that position. Bai et. al calculated the values for the free energies on different DNA sequences using the base pairing energies for the DNA bubble and the DNA-RNA hybrid [46]. These calculations (corroborated by fits to a seperate optical tweezers data set [4]), suggest that $\mathrm{TEC_{pre}}$ is energetically favored to $\mathrm{TEC_{post}}$, but that there is a wide range of $\Delta\Delta$G values based on DNA sequence (+1.6 +/- 1.8 $\mathrm{k_bT}$)

However, these calculations neglect other affects of DNA sequence that vary at each position, including the interactions between RNAP protein residues and specific DNA bases, as well as secondary structure forming in the growing RNA chain. A full energetic consideration for the free energy at each position is shown in Equation 2 [12]. The last two terms are more difficult to calculate, meaning that full determination of DNA sequence effect on free energy at each position must be experimentally determined.

$$\Delta G = \Delta G_{DNAbubble} + \Delta G_{RNA-DNAhybrid} + \Delta G_{RNAPbidning} + \Delta G_{RNA} \tag{2}$$

Evidence suggests that DNA sequence also affects the other reaction rates in the Brownian ratchet mechanism [46] [4]. Both the NTP binding kinetics and the rate of chemical catalysis depend on the specific NTP being incorporated (G,T,C,or U). While it is possible to directly determine the effect of NTP type on overall reaction velocity (ATP is fastest, UTP is slowest) [46] [4], calculating the NTP dependence of individual rate constants within each reaction cycle requires fitting to a specific kinetic model. As discussed in the previous section, these models are still being debated and often rely on unverified assumptions. Direct calculations of the NTP specific rate constants are still lacking.

0.1.4.4 Pauses during transcription

The pausing of RNAP serves many functions in transcription elongation. It helps synchronize trancription and translation [56], facilitates cotranscriptional folding of RNA [57] [58], and regulates trancription termination [59] [60] [61]. In addition, recent studies with human RNAP II suggest transcriptional pausing may play a role in human disease [62]. As discussed earlier, the first optical tweezers experiments of transcription elongation clearly identified RNAP pausing. Two types of pausing were detected: long-lived pauses (20s - 30 min) and short lived-pauses (or "ubiquitos" pauses, average of ∼3s) which accounted for ∼95% of pause events [63] [44] [64]. Over many DNA template positions, the frequency of pauses was found to be ∼0.9 per 100 bp, although slight variation in this value over different DNA template regions hinted at DNA sequence specificity in pausing [44]. Various models have been developed for RNAP pausing. In one view, originally developed by Galburt et. al, most if not all RNAP pausing events are the result of enzyme backtracking [65] [66][67] [68] [54]. The main evidence for this model comes from fitting a $t^{-2/3}$ power law to the measured pause durations (Figure 18) [65]. This time distribution is consistent with 2-dimensional diffusion driving a return from pausing after backtracking of the TEC. However, due to the spatiotemporal limits of most optical tweezers experiments, enzyme backtracking less than 3 bases during a pause has been impossible to directly detect.

In another view, pausing is again considered an "off-pathway" state, but not driven by backtracking. Instead, RNAP enters an "elemental pause state", or intermediate state that must be escaped to continue transcription [64] [44] [69] [70] [71]. Longer duration pauses could then arise from backtracking out of the elemental state [72]. Sturctural evidence for an elemental pause conformation has been demonstarted in bacteria [73]. A series of optical tweezers experiments out of the lab of Dr. Steven M. Block supported this idea of pausing [64] [44] [72]. Although exact DNA sequence registration between individual reads is difficult with optical tweezers (see optical tweezers background section 0.2.2), using a DNA template with the same 200 bp sequence repeated 8 times, Herbet et. al identified 6 DNA sequences

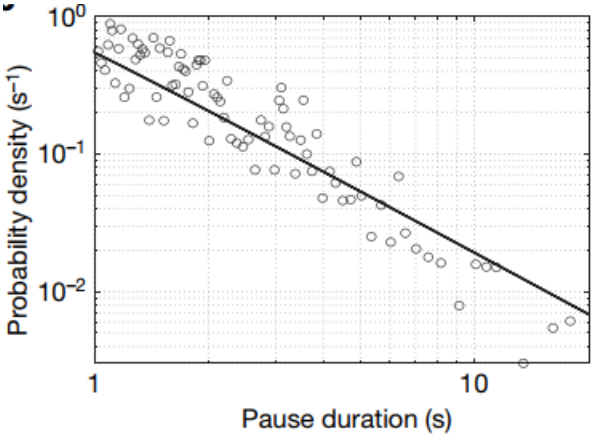


Figure 18: Pause durations determined by optical tweezers (Image reprinted from [65] with permission). The pause durations are fit by a $t^{-2/3}$ power law, supporting a random walk driving pause escape (backtracking model)

associated with pausing. At each pause location, a pause efficiency and pause lifetime were calculated (Figure 19). Pauses less than one second are difficult to detect with this method. Thus, exponential fits to the pause lifetimes greater than this threshold were used to calculate the number of short pauses (and a corrected total number of pauses). Pause efficiency (ϵ) was calculated by dividing the number of corrected pause events at a position by the total number of transcription reads over that position. They argue that the an "on-pathway" pause would necessitate a pause efficiency of 100%, and the value for every pause sequence is below that. Assuming an off-pathway mechanism for pausing, the reaction can be simply modeled as in Figure 20, where the reaction velocity is governed by the overall reaction rate k_f and the pause efficiency can be interpreted as in eEquation 3.

$$\epsilon = (k_p)/(k_p + k_n) \tag{3}$$

Further analysis and a subsequent optical tweezers study by Zhou et. al suggested that pausing at these locations was not caused by backtracking [64] [72]. By aligning and averaging over many instances of pausing at a particular sequence, an average TEC position for each pause sequence could be calculated. Comparing these positions to the specific nucleotide

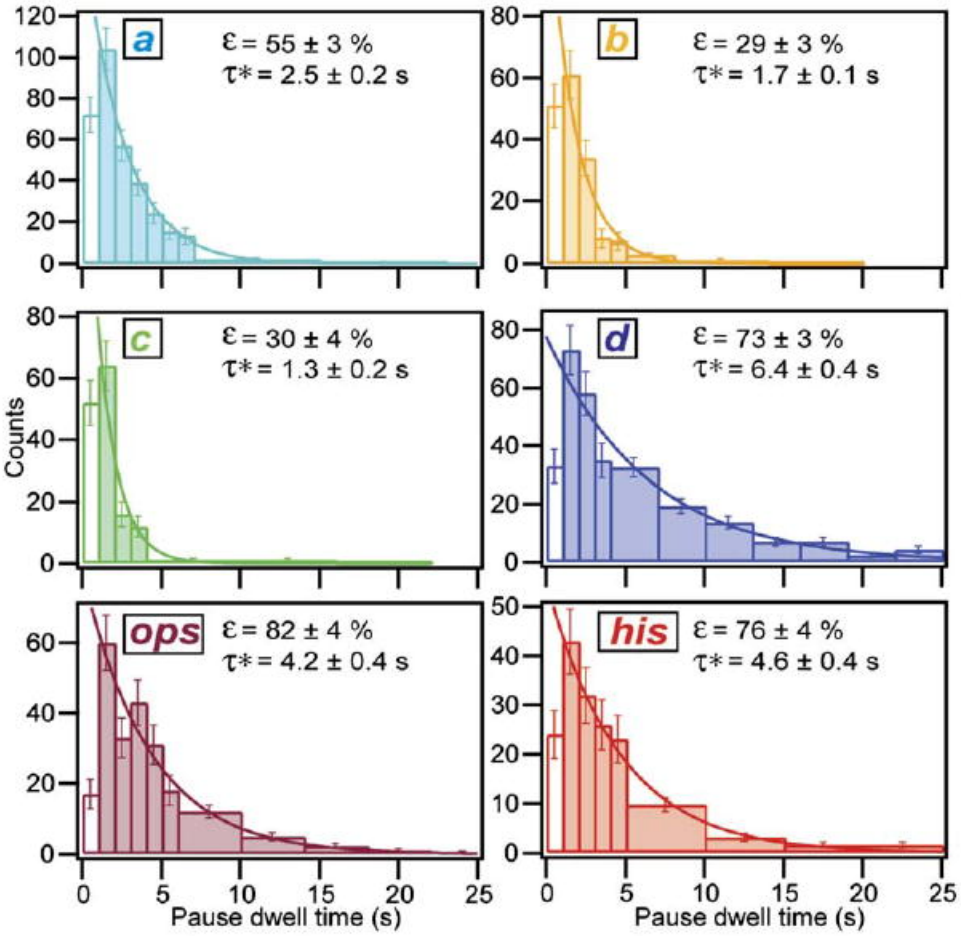


Figure 19: Pause lifetimes detected with optical tweezers: sequence specific pause sites (Reprinted with permission from [64]). Each panel depicts data from a different pause site along the scaffold (a,b,c,d, ops, and his) The pause durations at each site were fit by a single-exponential, excluding the first bin (below 1 second). The pause efficiencies (ϵ) at each site are below 100%, supporting an "off-pathway" mechanism for pausing

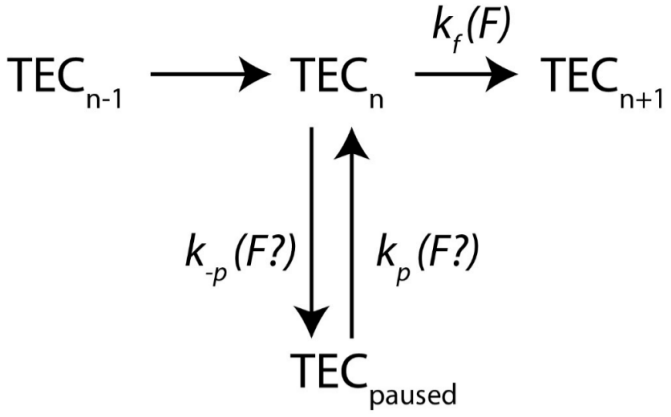


Figure 20: Simplified reaction model for RNAP Pausing. In a backtracking model for pausing, both the rate of pause entry k_p and pause escape k_{-p} depend on the applied force.

waiting to be incorporated for each pause (determined by biochemical gel assays), yields a pausing translocation state. According to this analysis, at most pause sequences, the enzyme is not backtracked, but instead has an average position somewhere in between TEC_{pre} and TEC_{post}. In addition, by varying the applied force and calculating the responding pause parameters, an elemental pasue was corroborated. The lifetime of a backtracked pause should depend on force; assisting forces will push the enzyme forward out of a backtrack. Pause duration was not affected by force, suggesting that only k_f) and not k_p or k_{-p} is affected by an applied force [72].

0.1.4.5 DNA Sequences that cause pausing: the elemental pause

These same experiments also determined a similarity between DNA template sequences responsible for pausing at different locations. More recently, bulk biochemical analyses have solidified this concept of commonality betrween pause sequences, fully mapping all of the genomic DNA sequences in *E. coli* that cause pausing. [74] [39]. Using a technique called NET-seq [75], Larson *et. al* identified 20,000 pause sites in *E. coli* [74]. Overlaying the DNA sequences of all pause sites can identify the RNA sequence that leads to pausing. The "consensus" pause sequence $G_{-10}Y_{-1}G_{+1}$, where Y refers to either a C or U, was corrobo-

rated by Vvedenskaya *et. al* Figure 21. While the bases in these positions (-10,-1,+1) are most important, surrounding sequence context also plays a role, as only ∼16 of sequences with a 1:1 match of $G_{-10}Y_{-1}G_{+1}$ actually lead to pausing [39]. Interestingly, this consensus pause sequence is satisfied by the Shine-Dalgarno sequence, the RNA sequence associated with translation start sites, further suggesting a role of sequence specific pausing in the coordination of transcription and translation [74].

The consensus pause sequence corresponds to unwinding of two G:C base pairs, one at the downstream stream edge of the transcription bubble (G +1) and another at the upstream end of the RNA/DNA hybrid (G -10). This alone should stabilize the pre-translocated state, as a G:C pair contains three hydrogen bonds (instead of two for A:T). However, the complementary pairs (ntDNA C at +1, or RNA C at -10) do not show up in the consensus sequence with any measurable frequency, suggesting that the energetic of DNA unwinding at these positions are not the dominant reason for pausing at the -10 and +1 positions. This suggests that pausing is the result of specific interactions between the RNAP protein and particular nucleobases at the downstream edge of the transcription bubble (+1 and -1) and the upstream end of the RNA:DNA hybrid)+10). In addition, it is interesting to note that there are no large-peaks in the logo plots for pausing) from the nucleotides upstream of position -10. This suggests that neither DNA duplex energies at the upstream edge of the transcription bubble or protein-DNA interactions in this region play a significant role in pausing. The nature of the protein- nucleic acid interactions at the positions associated with pausing are still undetermined in detail.

0.1.4.6 Structure of a paused RNAP

Until recently, X-Ray crystallography was the primary method for obtaining high-resolution structural data to compare different protein conformations over ngstrom distance scales. This method requires trapping an ensemble of molecules in a given state, so that crystallization occurs with a group of molecules in the same conformation. For RNAP, where the enzyme rapidly shifts between conformational states during transcription, obtaining an ensemble of

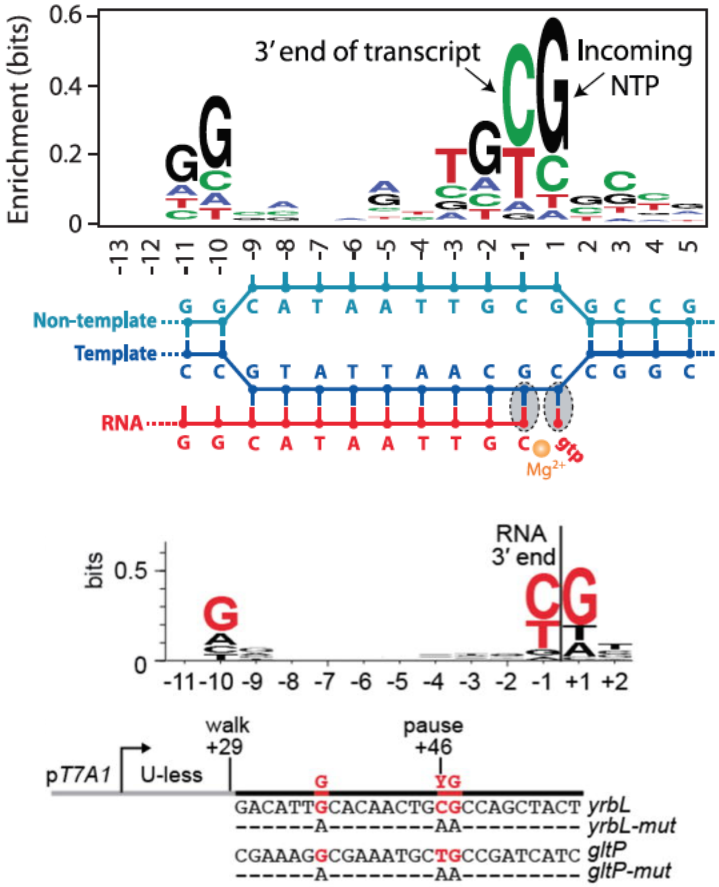


Figure 21: Consensus pause sequences (reprinted with permission from [74] and [76]). The DNA logo plots depict the non-template nucleotides associated with pausing at each position in the transcription bubble. Both studies determined a consensus pause sequence of $G_{-10}Y_{-1}G_{+1}$, where Y refers to either a C or U in the RNA. In this pause sequence, pause escape is achieved with the incorporation of a GTP at position 1. yrBL and gltP are specific DNA sequences that have been investigated and match the consensus pause sequence $G_{-10}Y_{-1}G_{+1}$ [76]

RNAP enzymes in the same state can be difficult. Nevertheless, methods to obtain crystal structures of RNAP have been developed and have helped tremendously in understanding transcription. For example, a crystal structure for *E. coli* RNAP in the Pre-translocated state was obtained using an RNAP initiation complex, where transcription has not yet begun but the enzyme happens to stall in a Pre-translocated register [77]. The development of a newer imaging method, cryo- electron microscopy (cryo-EM), where an individual molecular complex is imaged while frozen in-time, has provided access to structural data for many more RNAP states. A cryo-EM structure for the Post-Translocated state was obtained using a cross-linked *E. coli* RNAP [78]. Comparisons between these Pre [77] and Post [78] structures provide a clearer picture of typical RNAP translocation during transcription elongation, and reaffirm the basis for the translocation model presented in the previous section.

RNAP structures obtained during transcriptional pausing have suggested that a paused elongation complex (PEC) differs from both the Pre and Post structures. Specifically, a crystal structure of *Thermus* RNAP (another Prokaryotic, multi-subunit RNAP) obtained during pausing on a scaffold containing a pausing sequence showed an elongation translocation intermediate between Pre and Post [73]. In this intermediate state, the RNA translocates but the +1 tDNA is blocked from translocating by a kinked Bridge-Helix (BH), a helical protein domain that coordinates access to the active site. Without translocation of the +1 cytosine tDNA to the -1 position, the incoming GTP cannot match with the tDNA in the active site, providing a structural basis for transcription pausing. This is proposed to couple with large conformational changes of the protein, including widening of the RNA exit channel and an opening of the protein clamp

More recently, multiple cryo-EM studies have also revealed an intermediate state during pausing with *E. coli* RNAP at the his pause sequence [79] [80] . The his pause sequence contains the elemental pause sequence of $G_{-10}Y_{-1}G_{+1}$, but is preceded by an RNA hairpin sequence. Pausing at G+1, allows time for the RNA hairpin to fold, elongating pause lifetime by 10X. Guo et al. and Kang et al. obtained Cryo-EM structures during pausing, both before and after hairpin formation. The structures before hairpin formation represent

the structure of the elemental PEC (ePEC) common to all pauses at the elemental pause sequence. These structures were similar to the Thermus RNAP PEC, in that there is asymmetric translocation, where the +1 tDNA is blocked from entering the active site, while the RNA translocates. This results in a tilted RNA/tDNA hybrid, where some of the protein - nucleic acid interactions match with the Pre-Translocated state and some match the Post-Translocated state. The exact structure of this tDNA/RNA hybrid, particularly at the upstream end of the hybrid, varies between the two published cryo-EM ePEC structures, suggesting that more structures at even higher resolution need to be developed to determine the exact structure of the ePEC. These cryo-EM structures also predict small RNAP conformational changes when entering the ePEC, including a loosening of the swivel-module, regions of the B′ subunit of the protein including the clamp, dock, shelf, SI3, and the C-terminal segment. The rotation of this module was very small (about 5 degrees) prior to hairpin formation. The subsequent formation of the his hairpin in the RNA accompanies more dramatic conformational changes, including a large rotation of the swivel-module and a larger widening of the RNA exit channel

Additionally, intermediate states between Pre and Post with a tilted hybrid have been identified in structures of RNA polymerase II (Pol II) during initiation [81] and backtracking [81]. Structural data also suggests that the viral RNA-dependent RNAP enters an intermediate translocation state between Pre and Post during each reaction cycle of normal transcription elongation [82]. These results suggest that an intermediate state may be ubiquitous among RNAPs and could be part of the transcription pathway at more locations than just the elemental pause sequences.

Both papers detect a half-translocated state in between Pre and Post during initial pausing at this sequence. The idea is that the RNA and tDNA starts to translocate, but the tDNA is blocked by the Bridge-Helix (part of the RNAP) in the active site and what results is a 'titled' RNA-tDNA hybrid upstream of the active site. This is accompanied by rearrangement of a couple RNAP regions at the upstream end of the enzyme. There are even further structural changes once the hairpin forms and NusA binds. There are a few

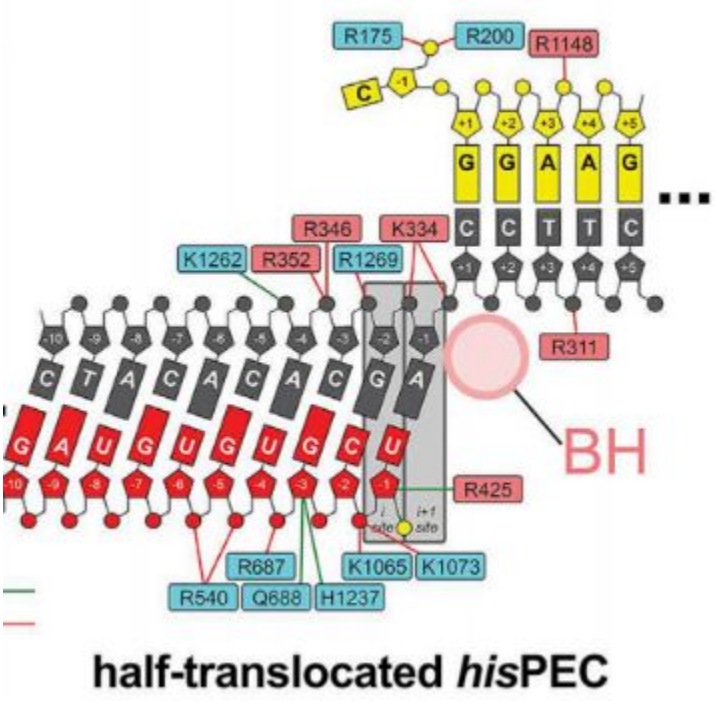


Figure 22: Half-translocated TEC detected using cryo-EM (reprinted with permission from [80]). At the *his* pause site, a TEC structure in between Pre and Post was detected. In this structure, the RNA:DNA hybrid are ′asymetrically′ translocated, where the tDNA resembles Post but RNA resembles Pre.

inconsistencies about details of this between the papers, but this overall idea is clear in both.

Structural information relating to a specific RNAP binding site involved in transcription initiation has shed light onto the mechanics of sequence specific pausing. The RNAP core recognition element (CRE) stabilizes a G at the +1 position in TEC_{post} [83] [76]. A mutant RNAP enzyme that loses an important residue in the CRE (RNAP D446A), shows increased pausing at $G_{-10}Y_{-1}G_{+1}$ sequences, suggesting that the CRE counteracts pausing at these sites [39]. Importantly, this supports the idea of pausing occurring prior to TEC_{post}, as the stabilization of G at +1 in TEC_{post} is required to escape a paused state. However, the exact mechanism of pausing, including the precise location of the enzyme throughout the duration of pause, is still up for debate.

0.1.4.7 Intrinsic termination of RNAP

During transcription termination, the TEC encounters a signal which instructs the RNAP to cease RNA elongation and dissociate from the DNA. Intrinsic terminators are DNA sequences that require no additional protein factors to induce termination. In prokaryotes, these are responsible for ∼50% of transcript termination. Typically, the DNA sequences are composed of a GC-rich palindromic region followed by at least four T-bases in a row, producing an RNA transcript with a stable hairpin structure followed by a U stretch [84] [85]. According to current understanding, the terminator sequence induces termination through the following mechanism: the TEC pauses along the T-stretch, allowing the RNA hairpin structure to fold [60] [86]. Next, RNA hairpin formation inactivates the TEC, displacing the RNA from the binding channel of the RNAP. This causes the TEC to rapidly dissociate [87] [60]. Differing models exist explaining the mechanistic role of the hairpin in this process. The forward translocation model [88] (Figure 22 A) suggests that the hairpin pushes the RNAP forward along the DNA template into a hypertranslocated (or forwardtracked) state that reduces the length of the DNA:RNA hybrid and destabilizes the TEC. Other models, including the allosteric model and the shearing model, propose that hairpin formation induces melting or shearing of the RNA:DNA hybrid without forward translocation [60] (Figure 23 B).

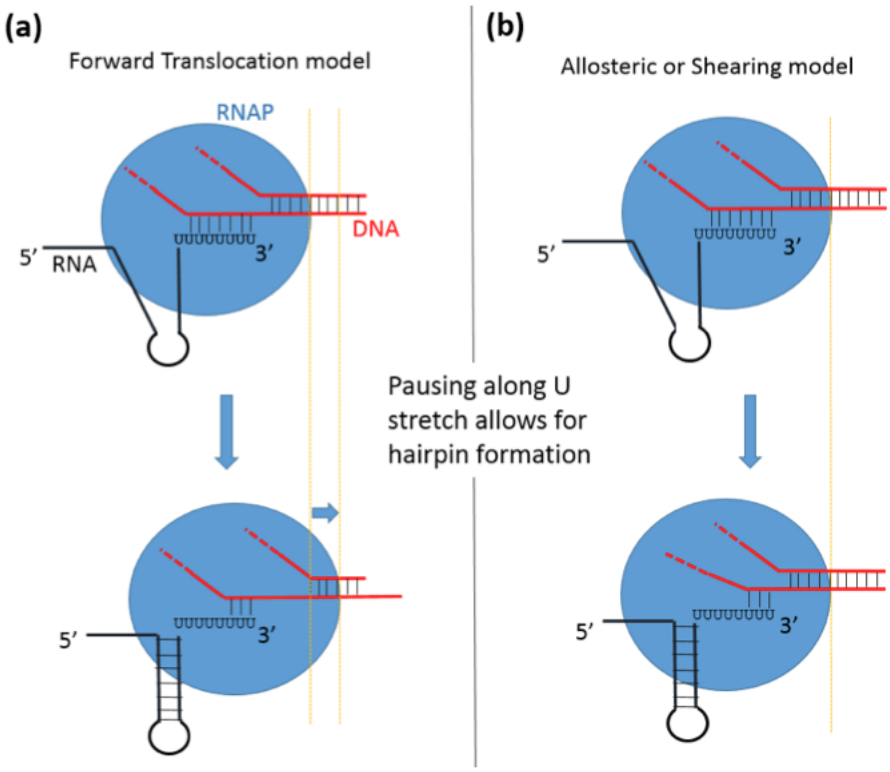


Figure 23: Models of intrinsic termination. In intrinsic termination, the enzyme pauses along the 'U-stretch,' allowing the formation of a hairpin in the RNA. This hairpin induces termination, either by promoting forward translocation of 4-5 bases and reducing the number of RNA:DNA contacts (as in A)), or by shearing the RNA:DNA base pairs without forward translocation of the TEC (as in B))

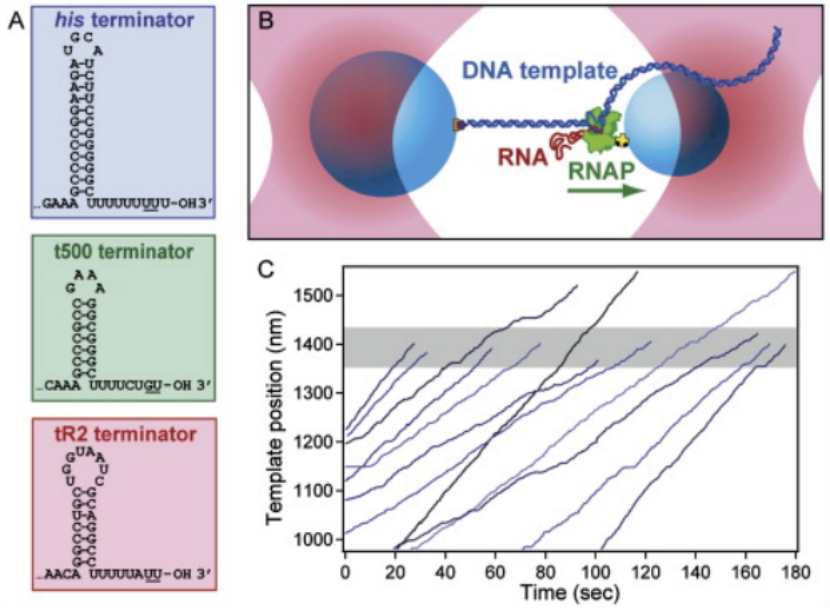


Figure 24: Optical tweezers experiments on intrinsic termination. Larson *et. al* [89] investigated intrinsic terminators (tr2,t600, and his) using optical tweezers. Relationships between termination efficiency and applied force supported the forward translocation model for t500 and a seperate mechnism for tr2 and his

Single-molecule technologies have been used to specifically investigate these models of transcription termination. Block and coworkers investigated the behavior of *E. coli* RNAP encountering intrinsic terminator sequences using an optical tweezers trapping assay [89]. In this study, Larson et al. investigated the effect of external load on termination efficiency (TE) at different terminator sequences (Figure 23). Two terminator sequences, his and tr2, exhibited negligible effects to TE under load, while a third terminator mutant, t500, displayed force-dependence. Because applied force should bias forward translocation, the researchers conclude that the forward translocation model fits with the t500 terminator, but an alternative process induces termination in his and tr2. This research suggests that multiple models for RNAP termination may be correct, depending on small differences in the terminator sequences. Although these findings are illuminating, direct observation of this proposed mechanistic difference is still lacking.

0.2 General Methods

The general experimental methods for nanopore sequencing with MspA and SPRNT have been overviewed elsewhere in this document (see background), in addition to published work [7] [8] [5]. In this section, I will detail the specific methods to apply SPRNT to *E. coli* RNAP core enzyme. The following section discusses the rationale for experimental methods and contains protocols that can be followed to design a SPRNT experiment on any DNA sequence. Specific methods pertaining to analysis used to obtain any of the results presented later in this document will be outlined at the end of each results subsections.

0.2.1 DNA Design for SPRNT-RNAP

Over the course of initial experiments, we determined a DNA/ RNA scaffold design that optimized:

1) Rate of RNAP capture into the nanopore (see section 4.1 for more details
2) ′Initiation′ of transcription after capture (see section 4.2 for more details)
3) adaptability to study many DNA sequences

The final DNA design contained 4 unique oligos (Figure 25) tDNA, ntDNA, RNA, and adapter DNA. The central 3 oligos (tDNA, ntDNA, RNA) form a DNA:DNA:RNA triplex, where the hydrogen bonding network linking the oligos was shared between the 3 strands. The important elements for experimental optimization are labeled in Figure 25)

Threading end: the 3' end of the tDNA contains a long single stranded poly-T region. This ssDNA tail is long enough to extend far away from theRNAP loaded onto the arrest sequence and can thread into the MspA nanopore. The extra Phosphate on the 3' end of tDNA adds a negative charge to the end of the DNA, further promoting DNA capture.

Adaptor: an adaptor DNA oligo is annealed to the 3' end of the ntDNA. The adaptor contains a duplex region, single-stranded poly T 5' tail, and a cholesterol attached to the 5' end. The cholesterol will embed into the lipid bilayer, localizing the RNAP/DNA complex to the region

Figure 25: Schematic of the nucleic acid scaffold used to investigate RNAP behavior. The scaffold consists of 4 unique oligos (3 DNA, 1 RNA). The important elements of this scaffold design (Threading end, arrest sequence, adaptor, variable section) are detailed in the main text of this section.

closer to the bilayer.

Arrest sequence:A particular sequence was designed to induce arrest of RNAP. This particular sequence tDNA:ntDNA:RNA (shown in red in Figure 25) causes the RNAP to load but then backtrack, so that the 3 end of the RNA sticks out of the NTP entry channel.

Variable section:The sequence downstream of the arrest sequence can be modified in order to watch transcription over various DNA sequence.

0.2.2 Protocols for SPRNT-RNAP

Preparation of DNA/RNA scaffolds: Oligonucleotides were mixed in the following ratio (1:1:2:5, tDNA:ntDNA:RNA:adapterDNA at 500nM tDNA) with 5X Annealing buffer (final 1X) and molecular grade H20. Mixture was heated to 95 C for 2 minutes; cooled to 25 C over a period of 70 minutes by decreasing temperature by 1 C per minute; stored at 4 C.

Annealing Buffer: 50 mM NaCl, 1 mM EDTA, 1 mM HEPES, pH 8.0

Preparation of arrested TECs: Annealed scaffold was mixed with RNAP core (1:10 DNA:RNAP with 50 nM with respect to tDNA in scaffold) with 10X Trancription Buffer (final 1X) and molecular grade H20. Mixture was incubated at 37 C for at least 30 mins before adding to SPRNT setup.

Transcription Buffer (1X): 100 mM KCl, 10 mM MgCl2, 1 mM DTT, 10 mM HEPES, pH

8.0

Nanopore isolation: A single M2-NNN MspA nanopore was established in a 1,2-di-O-phytanyl-sn-glycero-3-phosphocholine (DOPHPC) lipid bilayer using methods that have been well established (16). Lipids were ordered from Avanti Polar Lipids. MspA nanopores were isolated in a backwards configuration, with the smaller end of the nanopore facing the *cis* chamber. Backwards MspA was recognized due to distinct ion-current compared to forwards pore (∼230 pA backwards (vs. ∼180 pA forwards) at 180 mV with 500mM KCl in cis and trans at 21 C). Insertion probability between forward and backwards MspA was 50%. For detailed discussion on rationale for selecting backwards MspA and implications see *Determining ion-current patterns.*

Establishment of experimental buffer conditions: Final cis and trans experimental buffer conditions were established prior to addition of DNA/enzyme.

Cis Experimental Buffer: Transcription Buffer (1X) +[NTP] at specified concentration.

Trans Experimental Buffer: 500 mM KCl, 10mM HEPES pH 8.0

RNAP measurements: Arrested TECs were added to the cis chamber to a final concentration of 5 nM (with respect to tDNA). After a single arrested TEC was captured in the nanopore, returned from arrest, and transcribed the length of dsDNA, the TEC falls off the DNA scaffold and the DNA/TEC decouples from the nanopore. Multiple RNAP transcription events were therefore captured on each pore. Bulk activity of RNAP was prohibited due to ′arrest′ design so [NTP] was constant throughout experiment. NTPs were refreshed if experiment lasted more than 1.5 hrs.

0.2.3 Data Analysis Pipeline

Data acquisition: Data was acquired with custom labview software on an Axopatch 200B amplifier at 50 kHz. Ion-current and applied voltage data were isolated and downsampled by averaging to 5 kHz.

Data analysis: All of the data analysis was done using custom algorithms written in Matlab. The details of most of these algorithms (level-finding, alignment, consensus building) have been outlined elsewhere in published material. For the results presented in this document, the specific protocols used to obtain the results will be detailed in each subsection.

0.2.4 Materials

Proteins: MspA nanopore: M2-NNN MspA (accession number CAB56052.1) were prepared as described previously (15).

RNAP core enzyme: Obtained from Richard Ebright (Rutgers University). RNAP was stored at low glycerol concentrations (2%) and stored at -20 C in single use aliquots to avoid freeze/thaw cycles.

RNAP D446A: Obtained from Richard Ebright (Rutgers University).

Oligonucleotides: Oligos were purchased from Pan: Protein and Nucleic Acid Facility.

NTPs were purchased from Life Technologies.

0.3 Initial Results

The sections included under the category 'Initial Results' detail sets of reslts used to determine and optimize SPRNT experiments for RNA Polymerase. These initial results were used to justify the methods in this document as well as build the basis for understanding how to turn SPRNT positions measurements into meaningful kinetic measurements of RNAP enzymatic

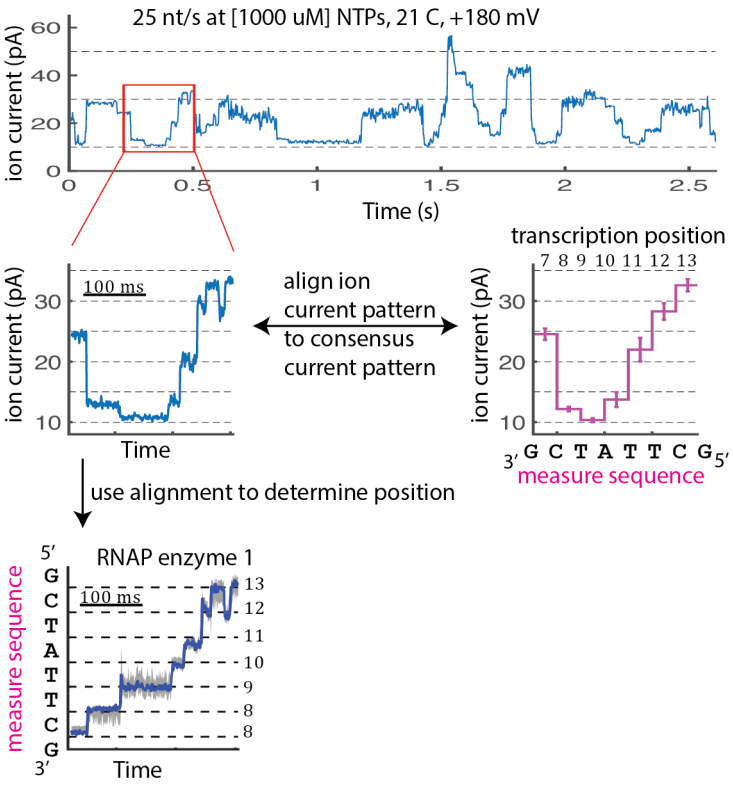


Figure 26: Ion current vs. time data is converted into DNA position vs time data. Ion-current data is recorded using the Axopatch 200B amplifier at 50 kHz and downsampled to 5 KHz (top panel). Discrete changes in ion-current correspond to individual steps of the enzyme threading the tDNA further into the pore with each step. The top panel depicts 50 discrete ion-current changes. Panel 2 depicts a smaller section of the same ion-current trace, where the measure sequence 3′ GCTATTCG 5′ (transcription positions 7 through 13) was threaded through the pore. The ion-current data is aligned to the consensus current pattern (panel 3) for the measure sequence at these positions. This alignment is used to generate the position vs time plot in panel 4. The same process was used to generate a position vs time trace for the entire read in panel 1. The details of this procedure are outlined in this section. While the position vs time plots were useful for visualizing the data, all of the analysis used in this study was done in ion-current space.

activity.

0.3.1 Determining ion-current patterns for 3′ threading on the backwards MspA pore

Introduction: The measured ion-current values for any DNA sequence in the nanopore depend on various experimental parameters including:
1) the orientation of the DNA with respect to the nanopore (5′ threading of DNA or 3′ threading of DNA) and

2) The orientation of the MspA nanopore with respect to the *cis* and *trans* sides of the membrane (forwards or backwards).

1) Because DNA is directional (has a 3′ and 5′ end), translocation of ssDNA through MspA can happen in two different orientations, referred to as 3′ first or 5′ first depending on which DNA end is entering the pore from the *cis* chamber. For SPRNT experiments with motor enzymes, both 5′ and 3′ first threading can be used to probe the enzyme in different manners depending on the experimental goals. For hel 308, a 3′ to 5′ helicase and translocase, 5′ threading will produce an assisting force while 3′ threading will produce an opposing force [9]. In addition, the motor enzyme - MspA interaction changes depending on the threading direction, as each direction means the opposite end of the enzyme contacts the pore. Certain motor enzyme - pore interactions may prevent proper enzyme function for SPRNT.

In 2015, our group published the mapping of measured ion currents through MspA for all of the 256 4-letter DNA sequences (e.g. AGTC) [32]). This map, a quadromer map, can be used to accurately predict the ion-current pattern for any sequence, and is required for any SPRNT experiment. The quadromer map provides a scaffold for constructing the consensus current plots used to generate the position vs. time plots in SPRNT. The published quadromer map was generated using 5′ threading of DNA. Earlier experiments using streptavidin to immobilize DNA in MspA showed that the measured currents for a given DNA sequence will change based on the threading direction [26] [25]. For example, cytosines have the deepest blockage current with 3′ threading compared to thymine with 5′ threading.

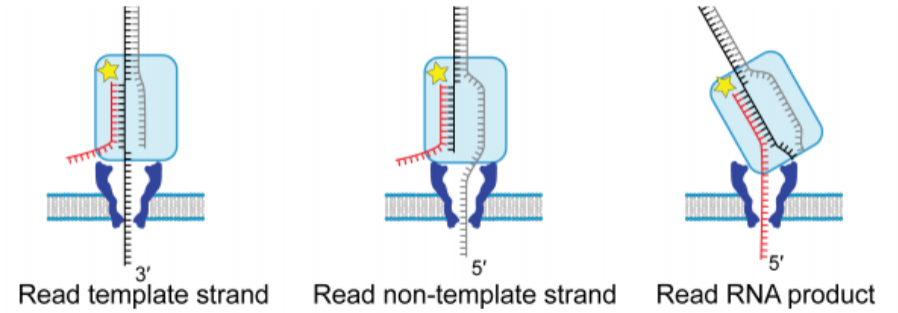


Figure 27: We experimented with three distinct threading designs for RNAP experiments with SPRNT. In all three schemes (read tDNA, ntDNA, or RNA) an assisting force was applied to the enzyme, but the interaction between RNAP and MspA changed. We successfully gathered data with both ntDNA and tDNA read, but had the highest throughput with tDNA, so used this orientation for all of the following experiments. For future SPRNT-RNAP experiments, a different orientation could be used to probe the enzyme in a different manner.)

This results in an entirely unique quadromer map for 3′ threading of DNA.

2) Backwards MspA (Fig 28), where the smaller end of the nanopore is on the *cis* side of the membrane, has a roughly equal insertion probability to forwards MspA (see general methods for identifying and isolating backwards pores). The backwards pore has a few experimental benefits compared to the forwards pore. The DNA capture rate into the pore is significantly greater on backwards MspA vs forwards MspA (unpublished), probably resulting from the decreased distance between rim and constriction on backwards MspA. This can improve experimental throughput. In addition, the MspA-motor enzyme interface in a SPRNT experiment will have a much smaller surface area on the backwards pore compared to the forwards pore. This means that the MspA-motor enzyme interaction may be less likely to inhibit enzyme function compared to the forwards orientation. Regardless of these potential benefits, all of the published nanopore experiments with MspA used the forwards orientation (the forwards orientation was assumed in the discussion above).

Early experiments with 5′ threading on backwards MspA immediately showed that the 5′

forwards pore quadromer map could not be used to accurately predict the ion current patterns for any DNA sequence with the backwards pore. The inverted orientation of the pore flips the DNA/ nanopore geometric alignment. The 5′ end of the DNA is now exiting the large side of MspA in trans. This results in a quadromer map qualitatively similar to 3′ threading on the forwards pore (e.g. cytosine is the deepest blockage with both 3′ threading forwards and 5′ threading backwards) but the details of the maps are unique as the direction of ion flow relative to the pore is different. This means that there are four unique quadromer maps, 5′ threading forwards, 3′ threading forwards, 5′ threading backwards, 3′ threading backwards. Two sets of the maps are similar (3′ threading forwards with 5′ threading backwards, AND 5′ threading forwards with 3′ threading backwards). Having an accurate quadromer map for each of the four orientations will be vital for the versatility of SPRNT moving forward.

Results: When we designed a nucleic acid scaffold to study RNAP with SPRNT, there were six candidate nucleic-acid ends to thread into the nanopore (3′ tDNA (upstream), 5′ tDNA (downstream), 3′ ntDNA (downstream), 5′ ntDNA (upstream) and 5′ RNA (upstream)). Any of these strands could have been promoted to thread into the pore by adding an extended single stranded tail with an additional phosphate at the end of the given strand (Figure 27). Threading of an upstream end will apply an assisting force to the enzyme while a downstream end, conversely, will apply an opposing force. RNAP has previously been shown to be much more sensitive to opposing forces, based off the much lower stalling force with optical tweezers assays [4]. In addition, the return from arrest methodology in these RNAP - SPRNT experiments required assisting force. For this reason, we chose to extend an upstream nucleic-acid end for threading into the nanopore in these experiments. Specifically, the tDNA was chosen as the extended strand for threading into the pore (instead of the ntDNA or RNA) for experimental considerations as initial tests (not shown) resulted in increased activity and return from arrest probability compared to the forwards pore. We speculate that this was due to a difference in the physical interaction between the enzyme and the nanopore on the forwards and backwards pore.

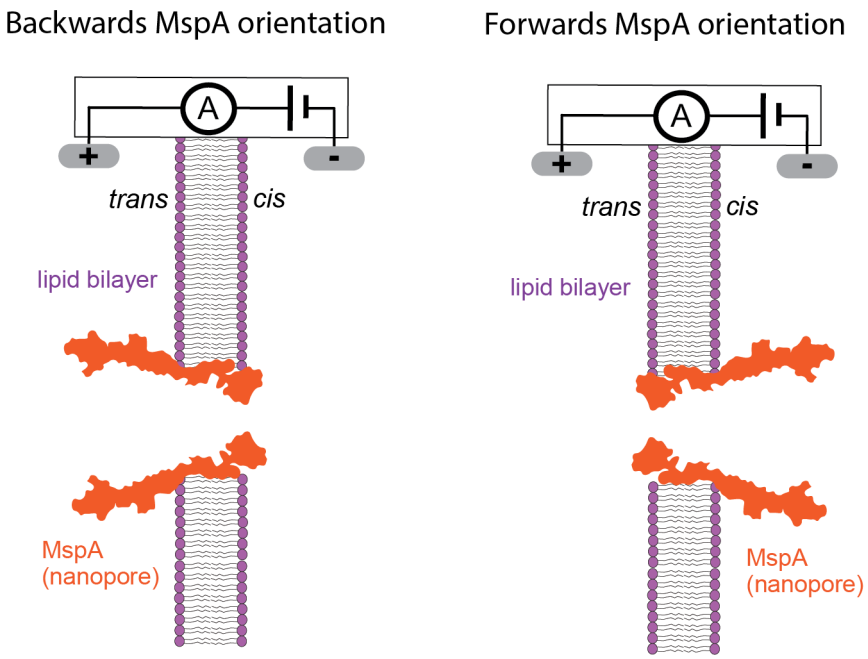


Figure 28: MspA inserts into the lipid bilayer in either a backwards (first panel) or forwards (second panel) orientation. The insertion probability was roughly equal for forwards vs. backwards MspA. All of the previously published MspA nanopore experiments have used the forwards orientation. The backwards pore offers advantages compared to the forwards pore (see text). For the RNAP - SPRNT measurements, we used the backwards orientation. For details on how to recognize and isolate backwards MspA, see general methods.

We also selected the backwards MspA conformation for this study, where the smaller end of MspA was on the *cis* side of the membrane (Figure 28). We chose this conformation over the forwards pore, as initial tests (not shown) resulted in increased activity and return from arrest probability compared to the forwards pore. We speculate that this was due to the difference in the physical interaction between the enzyme and the nanopore on the forwards and backwards pore.

These experimental choices, $3'$ threading on the backwards pore, required a unique quadromer map that has not previously been experimentally determined. To solve this, we used a separate enzyme that translocates on DNA to experimentally determine the measured ion-current patterns for each sequence. PcrAX is a mutant of an NTP-dependent DNA helicase that takes single-steps on DNA while translocating from $3'$ to $5'$ while unwinding DNA [90]. We measured PcrAX translocation with SPRNT with all the same sequences used in this study to determine the consensus ion-current patterns for each sequence with the backwards pore. The consensus ion-current pattern obtained with PcrAX matched well with the ion-current patterns obtained with RNAP (Figure 31) on its own. For this reason, we used the PcrA-derived consensus current patterns as an unbiased reference for mapping RNAP position during SPRNT with all of the sequences used in this study.

Methods: Consensus generation for SPRNT: the details of SPRNT consensus generation have been outlined elsewhere in published material [31]. For each DNA sequence used in this study, we gathered many reads with PcraX and performed both level finding and simultaneous multi-read alignment " by eye" in MATLAB.

0.3.2 Return from arrest and end of DNA scaffold

Introduction: Bulk enzyme activity can can make SPRNT experiments more difficult. Specifically, if RNAP continuously transcribes in bulk, the concentration of [NTPs] will change over the course of an experiment, and the location of the enzyme along the DNA scaffold will be unique for each complex captured by the nanopore. For nanopore experiments

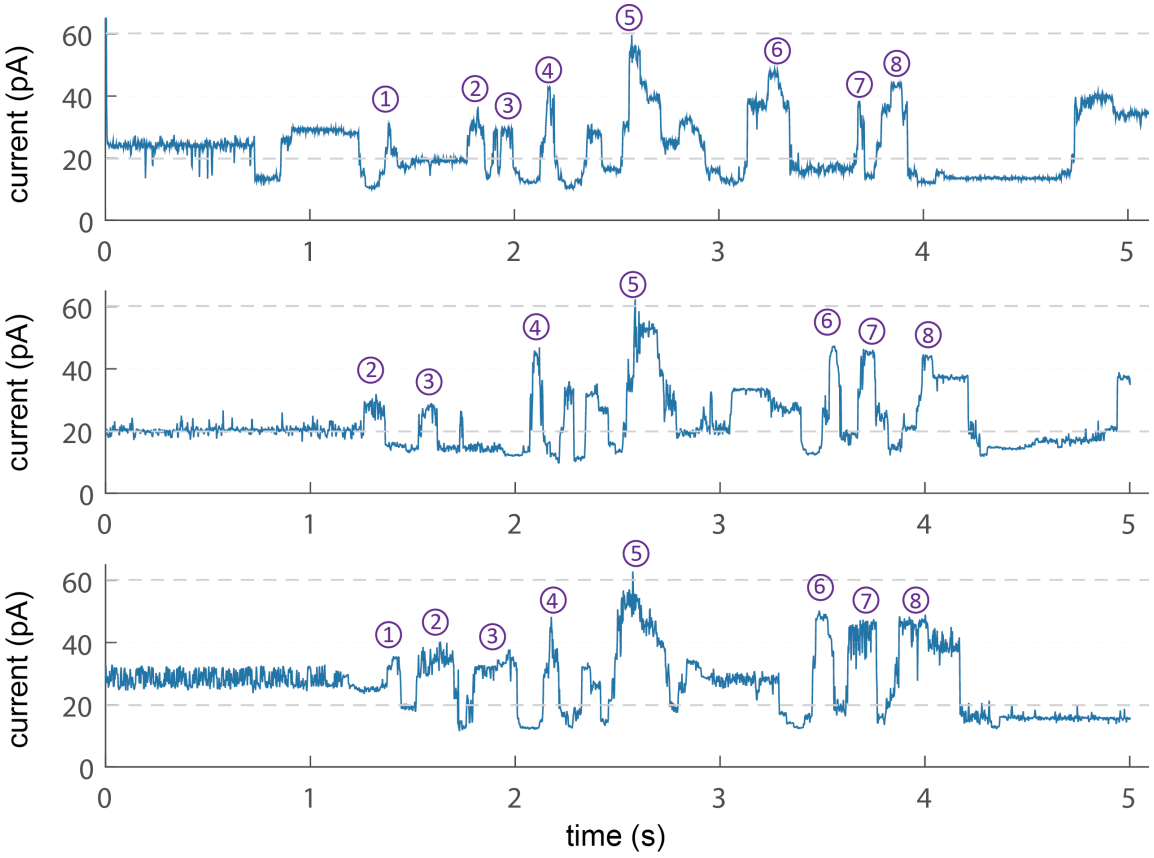


Figure 29: Three representative ion current traces of RNAP controlled DNA translocation through nanopore MspA. Although each trace corresponds to a unique RNAP molecule, the underlying current patterns are similar (as the DNA sequence is identical). Features of each trace are labeled 1 through 8, clearly marking the matching patterns.

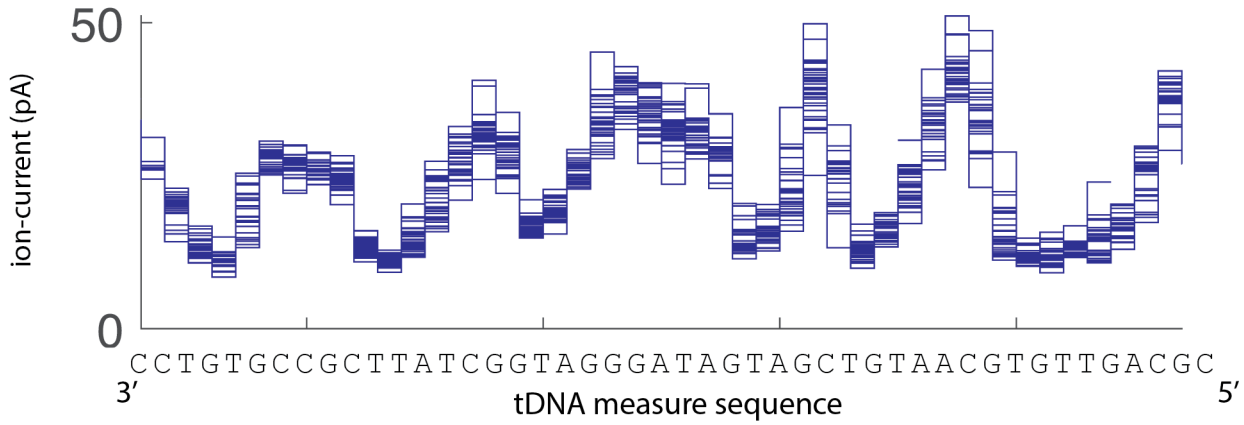


Figure 30: We gathered many reads with the same tDNA sequence and extracted the measure ion-current values for each read. This figure depicts the ion-current pattern for each read aligned with the tDNA sequence measured in the nanopore during each enzyme step. Although we were able to match the measured ion-current pattern to the DNA sequence (TT or TG produced the lowest currents while AA or AG produced the highest)

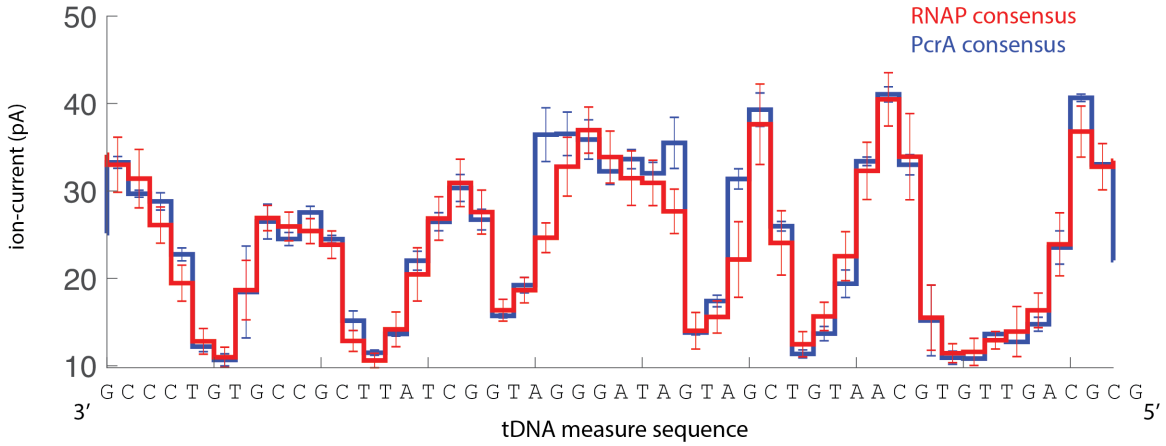


Figure 31: The consensus ion-current patterns for the tDNA measure sequence derived with PcrAX enzyme matched the ion-current patterns derived independently with RNAP. PcraX enzyme was used an unbiased reference map for all the DNA used in this study. Errors in ion-current are s.t.d.m

with phi29 DNAP, a "blocking oligo" was used to prevent bulk action of the polymerase. The force of the nanopore was used to strip off the blocking oligo, exposing a 3′ hydroxyl group for the phi29 DNAP to begin synthesizing DNA [7] (see background).

In vivo and *in vitro* RNAP can enter an arrested state, where the enzyme loses track of the 3′ end of the RNA by backtracking enough steps that the 3′ end of the RNA protrudes into the RNA exit channel. Even in the presence of high [NTPs] the enzyme stays arrested and transcription is inhibited. In bulk experiments, return from arrest has been demonstrated by the addition of protein factors GreB or GreA that induce transcript cleavage or by addition of protein Mfd, a translocase that pushes RNAP forward towards the 3 extendible end of the RNA [41] [91] [92].

We hypothesized that an arrested RNAP TEC (aTEC) could escape arrest by being pushed forward by the assisting force in SPRNT, acting in a similar manner to Mfd. By adding only pre-arrested TECs to the cis chamber, bulk activity would be limited and transcription would only occur after threading of the upstream 3 tDNA into the nanopore.

Results: Our collaborators at Rutgers experimentally determined a specific RNA:tDNA:ntDNA scaffold sequence that would load *E. coli* RNAP core, and without the presence of NTPs, could induce RNAP arrest at high efficiency (∼100%). No transcription was detected after addition of NTPs, but efficient (∼100%) return from arrest was detected after the addition of GreB. The specifics of the protocol for generating arrested complexes are detailed in the General Methods section.

We introduced an incubation of aTECs to a SPRNT experimental setup. We measured successful return from arrest judged by enzyme stepping events using this protocol over a range of applied voltages (120 mV to 220 mV) and a range of [NTPs] (10 μM to 1000 μM). We measured two distinct locations along the scaffold where the SPRNT positions measurements initiated. We call these locations Start Position 1 and Start Position 2 (Figure 34). SP 1 occurs upstream of the RNA 3 end. For measurements that began at SP1, the total time spent (T_{total}) in the transcription positions between SP1 and SP2 was [NTP] dependent

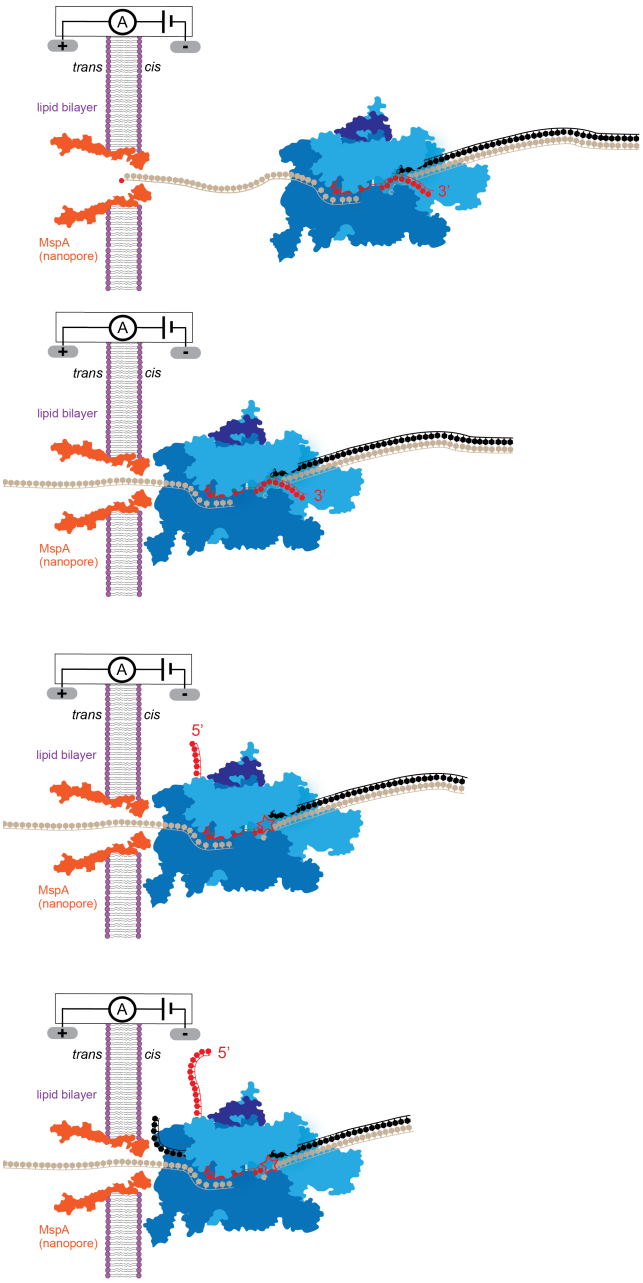


Figure 32: Return from arrest schematic with SPRNT: Arrested TEC, with an extended 3′ tDNA, is introduced to the cis chamber (Panel 1). The tDNA is pulled into the nanopore until the RNAP downstream edge contacts the pore (Panel 2). The force of the tDNA being pulled into the pore results in an equal and opposite force pushing the arrested TEC forward. the arrested TEC is pushed forward until the 3′ end of the RNA reaches the enzyme active site (Panel 3). Transcription begins and the tDNA is threaded further into the pore while the enzyme walks along the DNA, extending the RNA transcript (Panel 4)

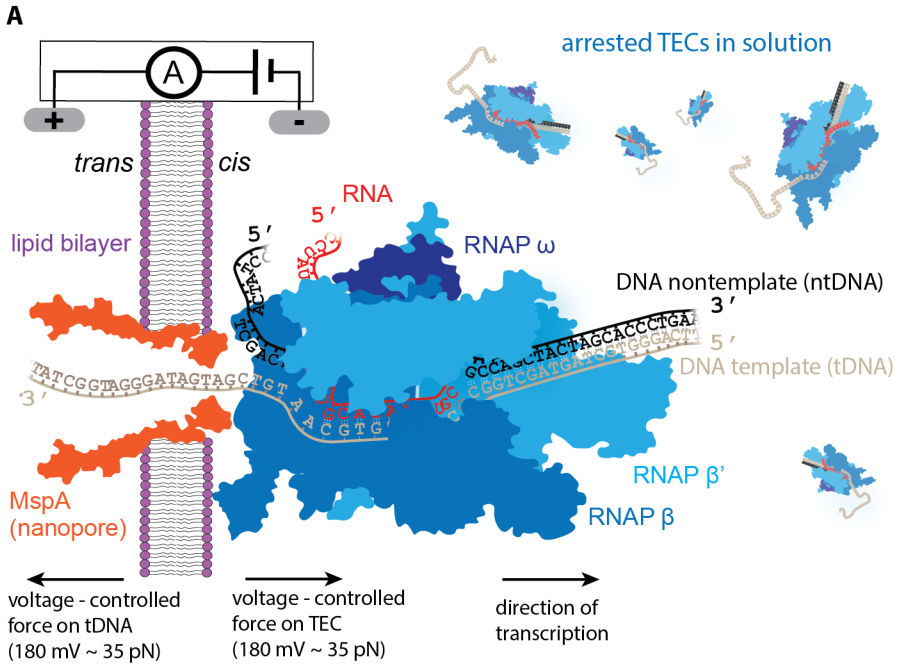


Figure 33: RNAP schematic during SPRNT: Close-up of Figure 32 Panel 4. The movement of the tDNA through the nanopore is controlled by RNAP during transcription. The applied voltage produces an assisting force on RNAP during transcription. The magnitude of the assisting force is proportional to the applied voltage. Arrested TECs in solution can enter the nanopore once the current RNAP finishes transcription and decouples from the nanopore.

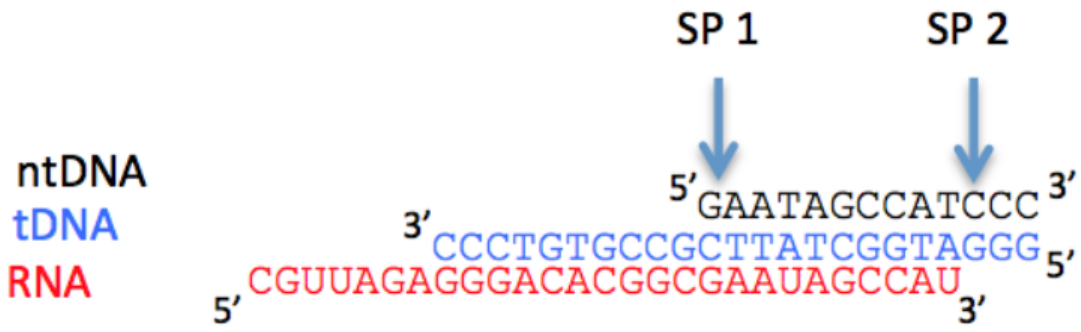


Figure 34: An arrest sequence for RNAP has been experimentally determined. The arrest scaffold consists of three oligos (ntDNA, tDNA, RNA) that are annealed (see general methods) to produce a nucleic acid triplex structure. RNAP loads onto this structure and slides backwards, so that the 3′ end of the RNA extends out the NTP entry channel. We detected transription start at two different locations (SP1 and SP2), suggesting two unique modes of RNAP return from arrest in SPRNT (see text)

(Figure 36). SP 2 occurs right near the 3 end of the RNA strand. For measurements that began at SP2, the total time spent (T_{total}, NTP = 1000 μM) at the two positions right after SP2 was much greater than at subsequent positions. We very rarely measured transcription starts far upstream (along the poly-t tail of the tDNA) and further downstream (past the arrest sequence).

For all of the enzyme measurements after return from arrest (regardless of start position), we also measured an increase in T_{total} at the yrbL pause sequence location (discussed elsewhere) and at the final five transcription positions before the end location (Figure 37).

Discussion: We have developed a method to successfully inhibit bulk activity of RNAP and initiate transcription by threading the upstream tDNA extended from an arrested TEC. This arrest sequence can be designed into any scaffold upstream of the DNA sequence of interest to be tested with SPRNT. Threading of the 3′ tDNA into the nanopore catalyzed return from arrest resulting in two unique start locations (SP1 and SP2), which correspond

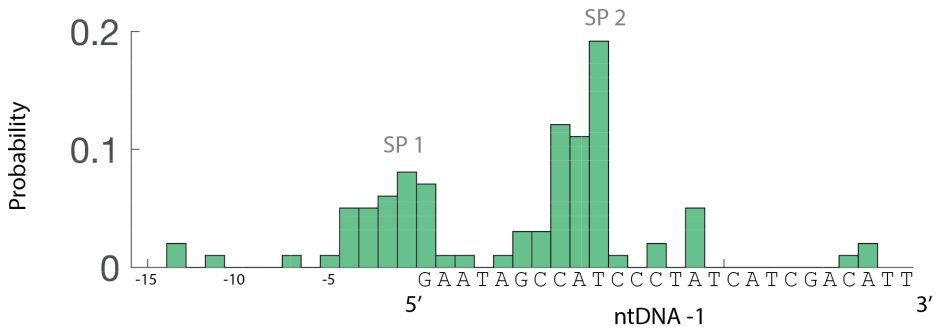


Figure 35: Histogram of RNAP start positions. We marked the first measured transcription position for every SPRNT read. There were two distinct regions where transcription start occurred. We defined the two peaks in the histogram as SP1 and SP2.

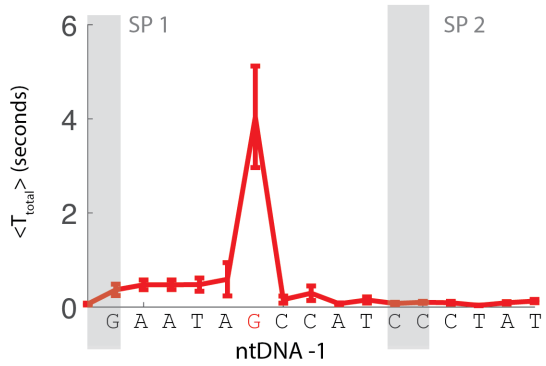


Figure 36: For reads that started at SP1, we calculated the T_{total} at each transcription positions with trace amounts of GTP ([GTP]= 0 μM, [CTP,UTP,ATP]= 1000 μM). We measured a spike in T_{total} at the G incorporation position in between SP1 and SP2. Data was acquired at 180 mV, 21C.

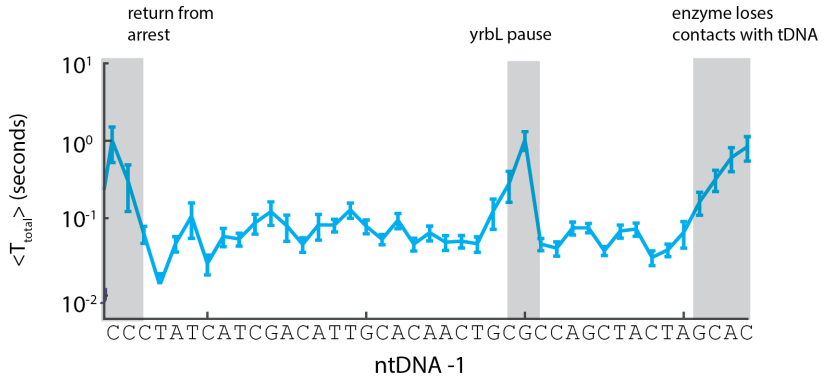


Figure 37: We tracked T_{total} across the DNA scaffold with [NTPs]= 1000 μM. The transcription positions assosicated with behavior outside of the normal transcription elongation pathway are marked in grey. Data was acquired at 180 mV, 21C

to two different mechanisms of return from arrest. Because the transcription rate between SP1 and SP2 was [NTP] dependent, return from arrest at SP1 was most likely the result of force-induced breaking of the RNA strand, exposing the 3′ end of the RNA in the active site at SP1. Transcription reads that began at SP2, on the other hand, were most likely the result of the enzyme quickly being pushed until the 3′ end of the RNA reaches the active site. Subsequently, the increase in duration at SP2 could be the result of misalignment or Hypertranslocation of the enzyme prior to the first NTP incorporation. Because of the different modes of return from arrest, in the subsequent results presented here, any measurements of transcription between SP1 and SP2 +3 were gathered using enzyme reads that began at SP1. All other data for the other transcription positions was combined between SP1 and SP2 reads. We also excluded the final four transcription positions from our analysis due to significant increase in total time spent in these positions. We believe this increase in duration was the result of the enzyme losing contacts with the downstream DNA as the double stranded scaffold begins to end.

0.3.3 Verification of transcription activity

Introduction: Because the SPRNT-RNAP measurements were performed at assisting force, we sought to determine whether the translocation of the tDNA through the nanopore in these experiments was the result of actual transcription activity by RNAP rather than force-driven unwinding of the DNA scaffold by RNAP without RNA synthesis (as in DNAP unzipping experiments with SPRNT [7].) Although single-molecule optical tweezers assays have previously been performed with *E. coli* RNAP at similar scales of assisting force [4], the application of force in these experiments was directed through a biotin tag on the B′ subunit and was therefore was not directly comparable to the force-application in SPRNT.

Results: We first performed SPRNT-RNAP experiments without adding NTPs (0 μM [NTPs], 180 mV, 21C) and did not detect any enzyme-controlled translocation of DNA through the nanopore. At the very least, this suggested that RNAP could not return from arrest without the presence of NTPs, but did not necessarily negate the existence of force-driven unwinding after arrest. Therefore, we reduced just the [GTP] to 0 μM (while keeping the other three NTPs at 1000 μM) and tracked many RNAP enzymes (n=20) at 180 mV and 21C. RNAP successfully returned from arrest in these conditions and proceeded to the first GTP incorporation site quickly (Figure 38). At the first GTP incorporation site, the enzyme stalled for long periods ($\sim$10 s) but would often continue forward returning to a quick rate until reaching the next GTP incorporation site and slowing again ($\sim$ 10 s). (The width of the peaks in $\langle$ T_{total}, GTP = 0 μM $\rangle$ / $\langle$ T_{total}, NTP = 1000 μM $\rangle$ across multiple transcription positions is further discussed later in this text).

Discussion: Force-driven unwinding without RNA synthesis would cause the enzyme to lose complete contact with the 3′ end of the RNA and therefore returning to synthesis necessitates either enzyme backtracking or reinitiation.

Therefore, because the enzyme traveled the transcription positions between the two GTP sites at a rate comparable to the rate at 1000 μM of all four NTPs, but then slowed again

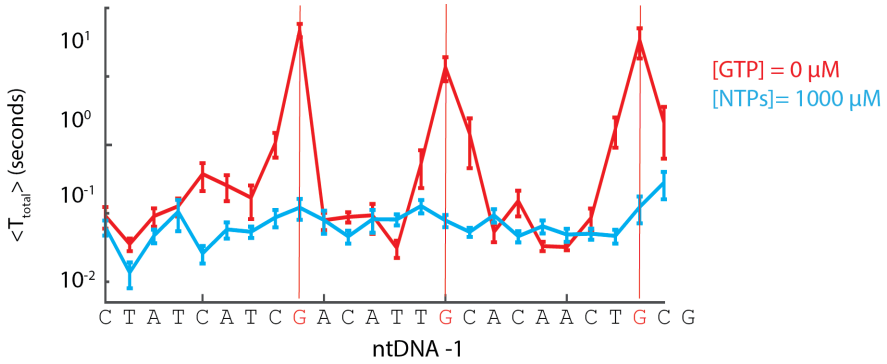


Figure 38: We tracked T_{total} across the DNA scaffold with [NTPs]= 1000 μM and with trace amounts of GTP ([GTP]= 0 μM, [CTP,UTP,ATP]= 1000 μM). T_{total} increased at positions associated with GTP incorporation. Data was acquired at 180 mV, 21C

at the next GTP site, we conclude that the progression of the enzyme past the first GTP incorporation site was the result of continued transcription and not force-driven unwinding. Either the incorporation at the first GTP site was the result of enzyme misincorporation (adding the wrong nucleotide C, U, or A) or the presence of very low concentrations of GTP in the other three nucleotide types, meaning the effective [GTP] was greater than 0 μM. Regardless, these results suggested that force-driven unwinding events in our study were not the cause of measured DNA translocation through the nanopore.

Methods: $\langle \mathbf{T}_{total} \rangle$**:** For every read, we calculated the total time at each position by summing all of the dwell times of every visit to that position (T_{total}). We calculated the average T_{total} ($\langle T_{total} \rangle$) by taking the mean of T_{total} across all reads. We calculated the error in $\langle T_{total} \rangle$ using the standard error in the mean (s.d.m.) $= \sigma/\sqrt{n}$

0.3.4 Implications of Varying Applied Voltage

Introduction: Although we can estimate the applied forces in SPRNT, a direct measurement of the applied force does not exist. This measurement will require coupling SPRNT with another single-molecule technique, like optical tweezers or magnetic tweezers where the force magnitude is inherent to the technique. Until this experiment is completed, we must

rely on estimates made from biophysical models of DNA stretching and measured stretch vs voltage relationships with SPRNT. The details of these estimates have been compiled elsewhere [5]. The voltage vs. force relationship calculated from these estimates can be used to estimate the force applied in these SPRNT-RNAP experiment (Figure 39. However, the error in these measurements is large. In addition, the estimates made with the forwards pore may not fully describe the force-voltage relationship with the backwards pore. Nevertheless, the voltage is proportional to the applied voltage, so a 50% reduction in voltage will result in a 50% reduction in applied force. Therefore, we can make SPRNT measurements at different voltages, and be confident in the relative changes in force across these various measurements.

In addition to reducing the applied force, lowering the applied voltage also reduces the flow of ions through the nanopore, which lowers the magnitude of the measured ion currents during DNA translocation through the nanopore [5]. Therefore, the change in measured ion current between consecutive measure sequences is reduced at low voltages. For this reason, it can be difficult to accurately identify enzyme steps at lower voltage. This limits full resolution SPRNT measurements to a specific range of applied voltages. In other words, for any enzyme studied with SPRNT, there is a minimum applied voltage (i. e. applied force) needed to maintain adequate resolution to detect enzyme state transitions. Applied voltage can also affect other experimental parameters, including DNA capture rate , and DNA phase shift [93], the stretching of DNA into the nanopore which is important for measuring ion current.

Results: We successfully gathered many SPRNT-RNAP reads with applied voltages from 120 mV to 220 mV. We attempted to gather data at lower voltages, but experimental throughput was greatly reduced below 120 mV, probably due to a combination of reduced tDNA threading and return from arrest probability after threading at lower voltages. The magnitude of measured ion currents was greatly reduced at 120 mV (compared to 220 mV) (Figure 40). For many of the measure sequences, identifying discrete enzyme steps was not possible at 120 mV due to the reduced signal-to-noise ratio. We did not attempt to acquire

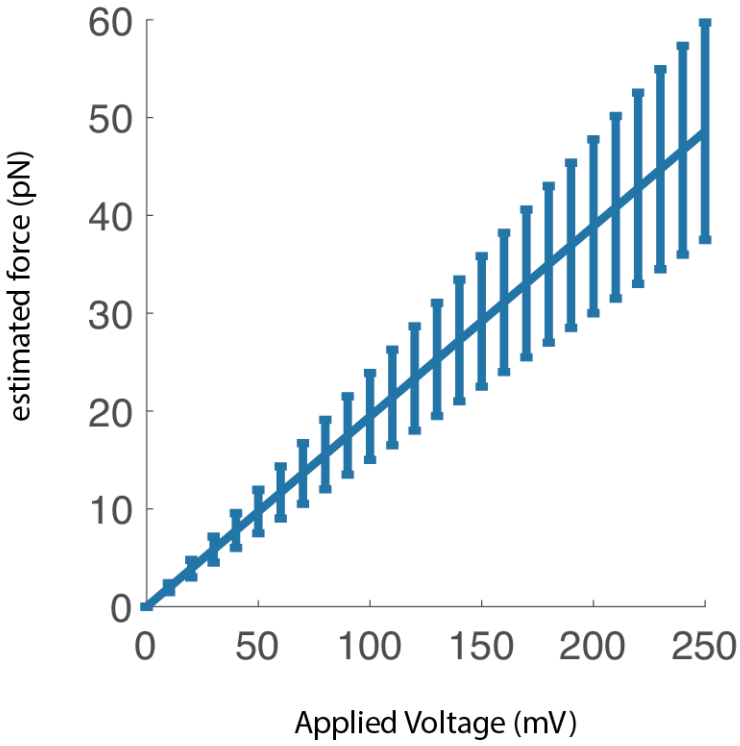


Figure 39: The relationship between force and applied voltage in SPRNT has been estimated but not directly measured. Although the rror in this estimate is large, applied force and applied voltage are directly proportional

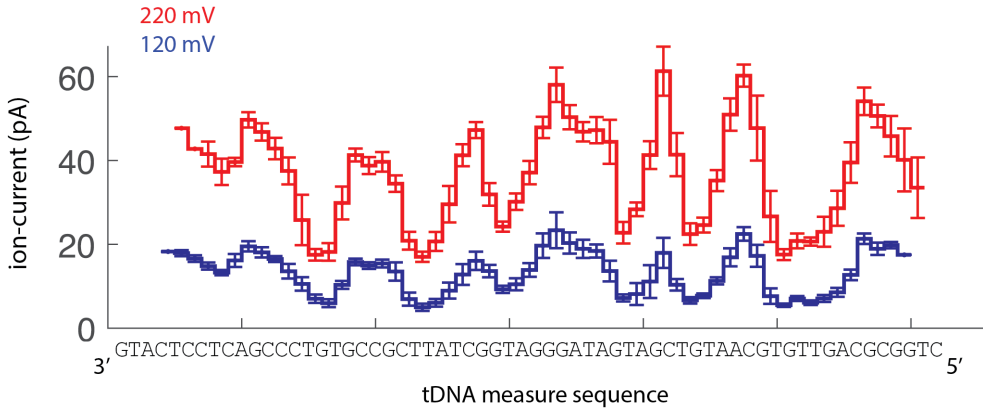


Figure 40: We determined the consensus ion-current patterns at various voltages. The magnitude of ion-current blockages is decreased significantly at lower voltage.

data at over 220 mV.

0.4 Results

0.4.1 Optimizing Reaction Conditions for SPRNT Experiments (Asymmetric Salts)

Note: This was my first major project in the nanopore lab. It was started during my time as an Amgen Scholar (REU program) and continued as a laboratory technician. The project was finished during my PhD tenure and published in PlosONE in 2017. This study was not conducted with RNA Polymerase, and is therefore a bit of a digression from the rest of the results presented in the following sections. However, the results are vitally important to choosing reaction conditions for any SPRNT experiment, including RNAP.

Publication: Nova, Ian C., et al. "Investigating asymmetric salt profiles for nanopore DNA sequencing with biological porin MspA." PloS one 12.7 (2017): e0181599.

Introduction: When investigating any nucleic-acid (NA) processing enzyme with SPRNT, operating conditions in the *cis* reaction chamber must match the preferred operating con-

ditions for the given enzyme. Many NA enzymes are particularly sensitive to high salt concentrations, as the electrostatic interactions governing the binding of nucleic-acids are salt dependent. Therefore, the preferred operating conditions for the majority of NA enzymes are [KCl] below 100 mM, mimicking the cellular environment of the organism from which the enzyme was found. However, operating in this regime of [KCl] poses a challenge with SPRNT, as the magnitude and noise of the ion current measurements are also [KCl] dependent. For this reason, the enzymes used for the initial nanopore sequencing experiments, hel 308 and phi29 DNAP, were both salt tolerant, working in conditions near 400 mM [KCl]. This provides adequate signal to noise ratio (SNR) for detecting enzyme steps with SPRNT.

We hypothesized that SNR could be maintained with low *cis* [KCl] by increasing *trans* [KCl], creating what we call an *asymmetric salt profile* (different *cis* and *trans* [KCl]). We tested this prediction by changing the concentration on each side of the membrane independently and monitoring corresponding changes in SNR when measuring DNA position. These asymetric salt profiles (low *cis* [KCl], high *trans* [KCl]) can be utilized to match the preferred operating conditions of any NA processing enzyme in SPRNT experiments.

Results:

Varying cis [KCl]

We varied *cis* [KCl] from 0 mM to 400 mM while keeping *trans* [KCl] at 500 mM and monitored phi29 DNAP-controlled DNA translocation through MspA nanopore (Figure 41). We generated a consensus ion current pattern of the same DNA sequence (Figure 41b) at each [KCl]. Using these consensus current patterns, we calculated the signal, defined as the average current difference between consecutive current states in the consensus, and the noise, defined as the average s.t.d.m. of the current states in the consensus at each [KCl]. The SNR increased over the range of [KCl] (Figure 41d), but the magnitude of increase across the whole range of *cis* [KCl] was small.

The unblocked MspA current (black dashed line in Figure 41A and Figure 41C) exhibits

an approximately linear response to *cis* [Cl-] at 500 mM *trans* KCl. Although the open pore current is significantly reduced at low *cis* [Cl-] (51.9 ± 1.2 pA at 20 mM *cis* [Cl-]), single MspA pores are still identifiable.

Figure 41A displays example current traces of phi29 DNAP controlling DNA translocation of the same DNA sequence in five separate *cis* conditions at 500 mM *trans* KCl. Consensus current traces (Figure 41B), generated from the set of translocation reads at each *cis* condition, demonstrate that the current pattern for the DNA sequence is consistent across salt concentrations even though the magnitude of each current level varies between conditions. For the DNA sequence analyzed here, the highest current level during DNA translocation involves an abasic site denoted by X. Both the highest (red dotted line in Fig 3A and 3C) and lowest (cyan dotted line in Figure 41A and Figure 41C) ion current levels, averaged over at least 8 events, increase with *cis* [Cl-]. The current range, the difference between the highest current level and the lowest current level, is a good metric to assess the signal magnitude at each asymmetric salt profile (magenta curve Figure 41C). The current range during DNA translocation increases only slightly with *cis* [Cl-] (Figure 41C). Between 20 mM and 420 mM *cis* [Cl-], the current range only increases by $\sim$ 30% while the open state current nearly doubles. At $\sim$70 mM *cis* [Cl-], the highest current during DNA blockage is equal to the open pore current. Below 70 mM *cis* [Cl-], the highest current level is larger than the unblocked pore current. Specifically, the highest current is 13.2 ± 1.6 pA greater than the open state at 20 mM *cis* [Cl-].

We next investigated the effect of *cis* [Cl-] on noise of translocation currents. For every DNA translocation event, we calculated the standard error for each current level between the lowest and highest currents (region shaded in grey in Figure 41B). The green stars in Figure 41D show the average for every noise measurement over the range of *cis* [Cl-]. Average current level noise is not statistically significantly correlated with *cis* [Cl-]. Between 20 mM and 420 mM [Cl-], the average noise remains relatively unchanged. With both signal and noise measurements available, we calculated the signal-to-noise ratio (SNR) for each *cis* [Cl-] using a t-test (Figure 41D). For every *cis* [Cl-] event, we calculated the average t-test value

for the current transitions during DNA translocation (again focusing on the shaded in grey in Fig 3B). The average SNR, plotted as the black circles in Figure 41D, follows a similar trend to the current signal range between 20 mM and 420 mM *cis* [Cl-], increasing from 3.7 $\pm$ 0.4 to 4.8 $\pm$ 0.5 ($\sim$ 30% increase). Decreased functionality of phi29 DNAP at high ionic concentrations prohibited experiments at *cis* [Cl-] above this range.

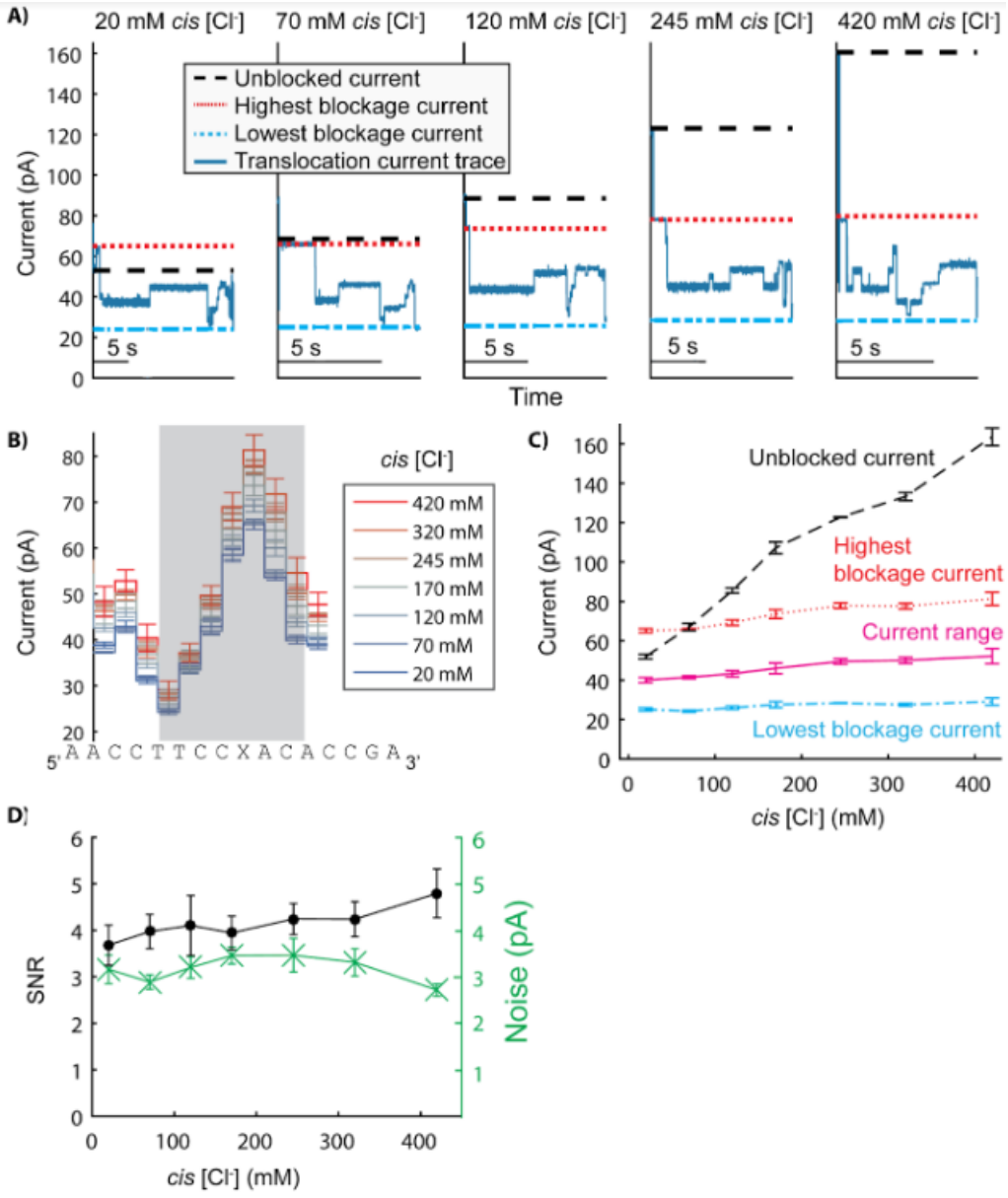

A)
20 mM cis [Cl-]
70 mM cis [Cl-]
120 mM cis [Cl-]
245 mM cis [Cl-]
420 mM cis [Cl-]
Unblocked current
Highest blockage current
Lowest blockage current
Translocation current trace
Current (pA)
5 s
Time
B)
cis [Cl-]
420 mM
320 mM
245 mM
170 mM
120 mM
70 mM
20 mM
Current (pA)
5' A A C C T T C C X A C A C C G A 3'
C)
Unblocked current
Highest blockage current
Current range
Lowest blockage current
Current (pA)
cis [Cl-] (mM)
D)
SNR
Noise (pA)
cis [Cl-] (mM)

Figure 41: (A) Current traces of phi29 DNAP controlled DNA *trans*location were recorded over a range of *cis* [Cl-] and at 500 mM *trans* [K+]. All experiments were performed at pH 8.00 ± 0.05. Panel A shows individual reads of the same DNA sequence over a range of concentrations (20 mM to 420 mM *cis* [Cl-] with an applied voltage of 180 mV). For each *cis* condition, the unblocked pore current (black dashed line) was calculated, as well as the highest current level (red dotted line) and lowest current level (cyan dotted line) using the consensus current traces from panel B. (B) At least 8 reads at each *cis* condition were used to generate a consensus current trace for the translocation of the DNA sequence in each condition. Panel B shows an 11 nucleotide section of the consensus plots for each condition. The 7 current levels shaded in grey were used for the SNR and noise analysis in panel D. Errors in current are S.E. The DNA sequence plotted underneath each current level correspond to the nucleotides in the constriction of MspA during that state. (C) Highest current level, lowest current level, and the range of current blockages were calculated for each *cis* condition using the consensus current traces from panel B. Errors are S.E. (D) Average noise and signal to noise ratio (SNR) were calculated for each *cis* concentration using only the level transitions in the region shaded in grey in panel B. Errors in average noise and SNR are S.E.M. (standard error of the mean).

Varying *trans* [KCl]:

We next varied *trans* [KCl] from 100 mM to 2 M while keeping the *cis* concentration of KCl at 200 mM with an applied voltage of 180 mV (Fig 42). We used a slightly different DNA sequence and gathered many phi29 DNAP controlled DNA translocation reads at each condition (N = 16 to 76). *Trans* [K+] ([KCl] + 8 mM) is indicated, as K+ is responsible for generating current from the *trans* chamber as the applied electric field direction drives K+ ions from *trans* to *cis*. Unblocked pore currents (black dashed line in Fig 42A and Fig 42C) vary significantly with *trans* [K+] at 200 mM *cis* KCl, although the relationship differs from the linear response seen with varying [*cis*] (Fig 42C vs Fig 42C). The DNA translocation current range drops off sharply at low *trans* [K+]. Between 108 mM and 158 mM *trans* [K+], the current range increases by almost 150 (12.9 ± 0.6 pA to 31.3 ± 0.8 pA). Small variations of *trans* [K+] produce large changes in signal over this range. However, as *trans* [K+] is increased, the magnitude of change in current range is diminished, exhibiting a plateau-like effect. Between 158 mM and 508 mM, the current range still increases by ∼ 40% (from 31.3 ± 0.8 pA to 44.9 ± 1.6 pA), but further increasing *trans* [K+] to 2 M only changes the

current range by $\sim$ 25% (from 44.9 $\pm$ 1.6 pA to 56.9 $\pm$ 3.3 pA).

We analyzed noise and SNR over the range of *trans* [K+] concentrations at 200 mM *cis* KCl, focusing on the level transitions in the region shaded in grey in Fig 42B. Average DNA translocation current noise (green stars in Fig 42D) varies more significantly for increasing *trans* [K+] than for increasing *cis* [Cl-] (Fig 42D vs Fig 42D). While current range changes dramatically between 100 mM and 200 mM *trans* [K+] (Fig 41C), the majority of the noise increase occurs above this range, between 200 mM and 1 M *trans* [K+]. Correspondingly, the average SNR, plotted as the black circles in Fig 42D, rises sharply between 108 mM and 158 mM, increasing by a factor of $\sim$3 (1.5 $\pm$ 0.3 to 4.7 $\pm$ 0.6). As *trans* [K+] is further increased, the SNR begins to decrease due to the increased noise and plateauing signal magnitude. At a fixed *cis [KCl]* of 200 mM, increasing the *trans* [K+] above 200 mM provides little benefit in distinguishing current transitions.

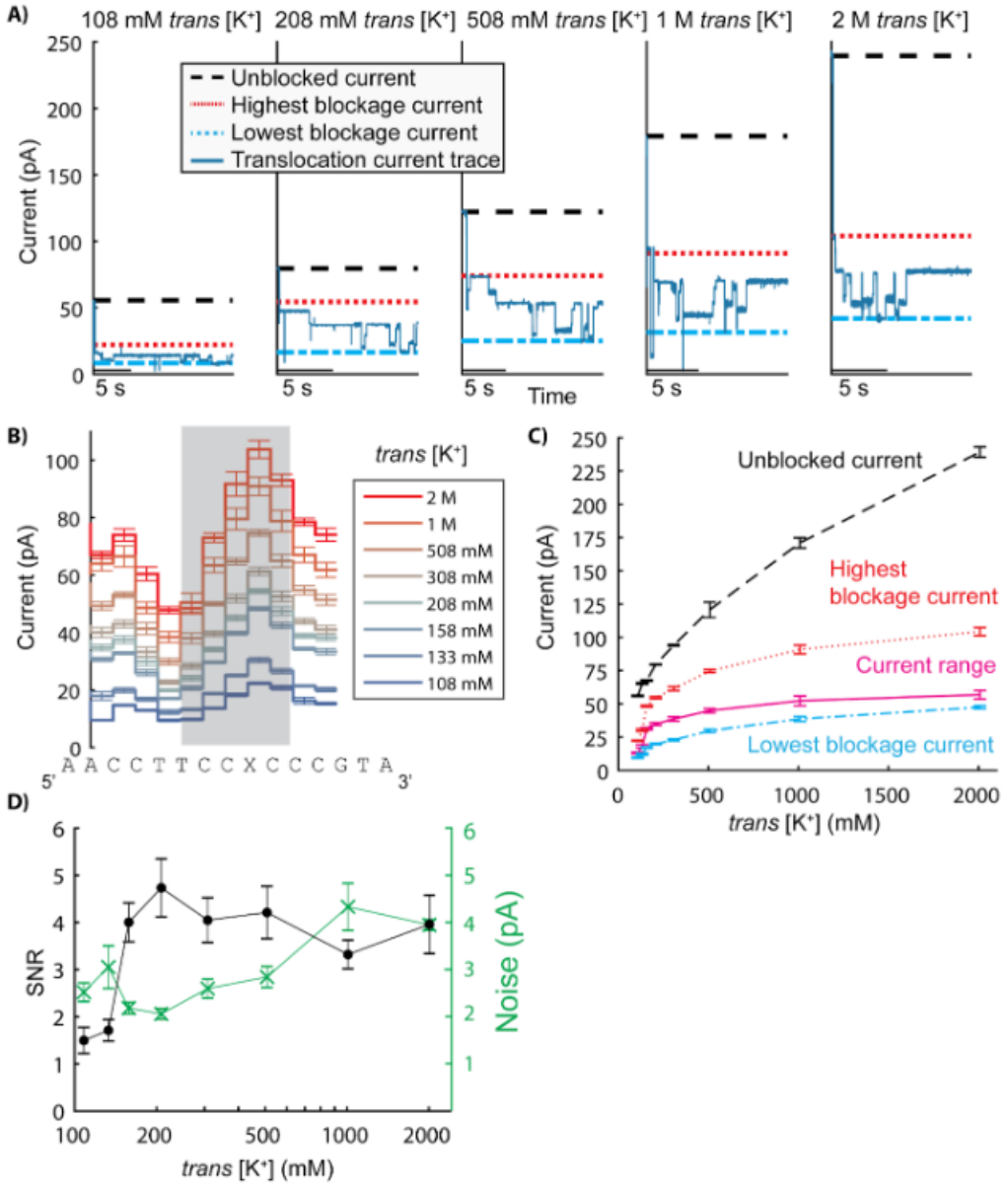

A)
108 mM trans [K+]
208 mM trans [K+]
508 mM trans [K+]
1 M trans [K+]
2 M trans [K+]
Unblocked current
Highest blockage current
Lowest blockage current
Translocation current trace
Current (pA)
5 s
Time
B)
trans [K+]
2 M
1 M
508 mM
308 mM
208 mM
158 mM
133 mM
108 mM
Current (pA)
5' A A C C T T C C X C C C G T A 3'
C)
Unblocked current
Highest blockage current
Current range
Lowest blockage current
Current (pA)
trans [K+] (mM)
D)
SNR
Noise (pA)
trans [K+] (mM)

Figure 42: (A) Current traces of phi29 DNAP controlled DNA translocation were recorded over a range of *trans* [K+] and at 200 mM *cis* [Cl-]. All experiments were performed at pH 8.00 ± 0.05. Panel A shows individual reads of the same DNA sequence over a range of concentrations (108 mM to 2 M *trans* [K+] with an applied voltage of 180 mV). For each *trans* condition, the unblocked pore current (black dashed line) was calculated, as well as the highest blockage current (red dotted line) and lowest blockage current (cyan dotted line) using the consensus current traces from panel B. (B) Multiple reads at each *trans* condition were used to generate a consensus current trace for the translocation of the DNA sequence in each condition. Panel B shows an 11 nucleotide section of the consensus plots for each condition. The DNA sequence plotted underneath each current level correspond to the nucleotides in the constriction of MspA during that state. The 5 current levels shaded in grey were used for the SNR and noise analysis in panel D. Errors are S.E. (C) Highest current blockage, lowest current blockage, and the range of current blockages were calculated for each *trans* condition using the consensus current traces from panel B. Errors are S.E. (D) Average noise and signal to noise ratio (SNR) were calculated for each *cis* concentration using only the level transitions in the region shaded in grey in panel B. The x-axis is logarithmic. Errors are S.E.M.

Discussion:

Enhanced K+ selectivity during DNA translocation: This study helps uncover the individual contribution of *cis* and *trans* ions to the ion current through MspA, revealing that the presence of negatively charged DNA in the pore promotes the transit of cations from *trans* to *cis* and minimizes the transit of anions from *cis* to *trans*. While the applied electric field controls the direction of ion flow (positive ions (K+) from *trans* to *cis* and negative ions (Cl-) from *cis* to *trans*), the rate of ion passage in each direction is also dependent on the microenvironment within the pore that the ions must pass through. The MspA M2NNN mutant used in this study is neutral within the pore constriction, with neutral asparagines replacing negatively charged aspartic acid residues in the wild type [4]. During DNA translocation, the constriction environment changes both physically and electrostatically. Along with the DNA bases occluding the pore constriction, each phosphate along the backbone of the DNA introduces a negative charge into the pore. By comparing the ion current through MspA with and without DNA (blocked pore current and open state current Figs 41 and 42) in identical ion concentrations and applied voltage, the overall effect of this altered microenvironment

on ion passage becomes apparent. In all of the asymmetric salt profiles sampled (Figs 41 and 42), except below ∼70 mM *cis* at 500 mM *trans* [KCl], the open state current is larger than the highest DNA blockage current. The combined rates of Cl- flow from *cis* to *trans* and K+ flow from *trans* to *cis* are reduced by the presence of DNA within the pore.

Assessing the relationship between ion current signal and *cis* and *trans* ion concentration independently can untangle the specific contribution of each ion type to the overall ion current. The relationship between ion current and *cis* [Cl-] (Fig 41C) demonstrates that the presence of DNA diminishes the ion current sensitivity to *cis* [Cl-]. As *cis* [Cl-] varies from 20 mM to 420 mM with a constant *trans* [KCl] of 500 mM, the currents through MspA with DNA present in the pore (highest and lowest blockage currents) increase at a much slower rate over the entire range than without DNA (open state current). However, with minimal Cl- present to flow from *cis* to *trans* (below ∼70 mM *cis* Cl-), the highest DNA blockage current actually exceeds the open pore current, indicating that while Cl- flow from *cis* to *trans* is minimized by the presence of DNA within the pore, K+ flow from *trans* to *cis* is enhanced. The DNA sequence leading to the highest current level in this strand contains an abasic residue, consisting of a missing base and only DNA phosphate sugar backbone, further suggesting that the negative charge promotes cation passage instead of some DNA-nucleobase-specific interaction with the cations.

The relationship between *trans* [K+] and DNA blockage current is consistent with this model of ion flow (Fig 42). Below ∼200 mM *trans* K+ at 200 mM *cis* KCl, the DNA blockage currents drop off steeply (Fig 42C) to the point at which distinguishing adjacent current translocation steps becomes prohibitively difficult below 100 mM *trans* K+ (Fig 42B and Fig 42D). This extreme sensitivity to *trans* [K+] lessens as *trans* [K+] is increased above 200 mM, and the DNA blockage currents begin to saturate. Current saturation is consistent with other biological porins that demonstrate large selectivity bias (preferring transit of one ion species over another) due to charged residues [15, 16]. In these previous studies, the proposed model suggests that each negative charge within the pore represents a binding site for a positive ion during transit. At a high positive ion concentration, all of these ion binding

sites are always occupied and mask the negative charges, which discourages further K+ ions moving from *trans* to *cis*. The limiting rate for ion transfer is no longer the time it takes to fill an evacuated site, but instead the time it takes for the positive ions to move from one site to the next. At this point, the flow of ions is independent of ion concentration and depends only on the overall membrane potential or applied voltage.

Overall, these phenomena suggest a complex model of ion flow for DNA translocation currents. When DNA enters the pore, the cationic selectivity of MspA increases as the charged backbone actually promotes K+ flow from *trans* to *cis* and minimizes Cl- flow from *cis* to *trans*. The specific nucleobases further modulate and block this flow depending on their type as the motor enzyme pulls DNA through the pore.

Benefits of asymmetric salt profiles for nanopore sequencing:

Salt concentration affects many experimental parameters in nanopore sequencing, including signal-to-noise ratio (SNR) during DNA translocation, motor enzyme function, and DNA capture rate. Using the results presented in this study, *cis* and *trans* conditions can be tuned independently to balance and optimize these parameters. Increasing SNR improves resolution when monitoring DNA translocation and could lead to improved sequencing accuracies. Due to the sharp dropoff in SNR at low *trans* [K+] (Fig 42D), resolving DNA translocation becomes difficult at *trans* concentrations below 150 mM. Additionally, increasing *trans* [K+] above 200 mM does not improve SNR, suggesting an ideal *trans* [K+] of above ∼150 mM to 200 mM with an applied voltage of 180 mV. Increasing *cis* [Cl-] also marginally improves SNR (Fig 41D), although even at negligible *cis* [Cl-], as long as *trans* [K+] is maintained, the SNR is not prohibitive to monitoring DNA translocation.

Simultaneously, altering the *cis* well conditions can affect the activity of the motor enzyme controlling DNA translocation in nanopore experiments. Salt concentration can affect the stepping behaviour, processivity, and binding affinity of motor enzymes [17, 18]. Even with reasonable SNR, enzymatic missteps along DNA (backtracking, skipping, or toggling between translocation states) obfuscate analysis and contribute to errors in nanopore DNA

sequencing. Because the motor enzyme is only present in the *cis* chamber and not *trans*, by tuning the *cis* salt concentration independently to the preferred operating conditions of the motor enzyme, enzymatic missteps may be minimized. Phi29 DNAP, the motor enzyme used in this study, can operate over a wide range of salt concentrations, enabling nanopore DNA translocation experiments with negligible *cis* [KCl] and up to 400 mM KCl (Fig 41). However, many motor enzymes that may be useful for nanopore sequencing (polymerases, helicases and recombinases) function only in a narrower range of salt concentrations, and many lose binding affinity above 50 to 100 mM (19,20). By maintaining the *trans* concentration for SNR, and adjusting the *cis* concentration specifically for enzymatic function, salt sensitive motor enzymes can be used for nanopore sequencing.

Additionally, both *cis* and *trans* concentrations can be altered to affect experimental throughput. In experiments with solid-state nanopores, Wanunu et al. demonstrated that asymmetric salt profiles with $[trans] > [cis]$ enhance the electric field in *cis* and subsequently increase DNA capture rate [11, 19]. While this same phenomena has also been observed in biological nanopores, a separate effect simultaneously influences DNA capture rate with changing *cis* concentration due to the electrostatic interactions between a biological pore and DNA. Specifically, Jeon et al. determined that increasing *cis* ion concentration shields the electrostatic interactions between the biological pore -hemolysin and a negatively charged polymer [9]. With -hemolysin at pH 7.5, these interactions are repulsive, and, therefore, increased ionic shielding promotes polymer capture. In our experiments with MspA, with three separate *trans* KCl concentrations, we see an increased DNA capture rate as *cis* [KCl] is increased at pH 8 (Fig 43). Although MspA mutant M2NNN is neutral within the constriction, the rim of the pore is net-negatively charged, creating an electrostatic barrier for DNA capture similar to -hemolysin at the same pH. While the benefit of increasing *cis* KCl concentration is apparent at all of the asymmetric salt profiles tested, increasing *trans* concentration at a given *cis* concentration also promotes DNA capture, suggesting that the ratio of *trans* to *cis* concentration still influences DNA capture with MspA.

When choosing *cis* and *trans* concentrations for nanopore sequencing with MspA, op-

timizing DNA capture rate can decrease the requisite amount of initial reagents (DNA, enzymes) to maintain experimental throughput. As described above, *cis* concentrations are limited by the operating conditions of the motor enzyme controlling DNA translocation. However, even though the SNR benefits of increasing *trans* concentration are negligible above ∼150 to 200 mM (Fig 42D), *trans* concentrations can still be maximized to increase DNA capture rate.

In summary, using asymmetric salt profiles, i.e. tuning *cis* and *trans* salt concentrations independently, allows simultaneous optimization of multiple parameters for nanopore DNA sequencing with MspA. *cis* concentration can match the preferred operating conditions of the motor enzyme controlling DNA translocation, even for salt-sensitive motor enzymes, with minimal effect to SNR. High *trans* concentration optimizes SNR and experimental throughput. By following these guidelines, new DNA motor enzymes can be effectively tested for nanopore DNA sequencing with MspA. Furthermore, in addition to DNA sequencing applications, the data gathered with each new DNA motor enzyme provides valuable scientific insight into enzyme function. By monitoring the controlled translocation of DNA through MspA by an enzyme, precise information about the kinetics and stepping behavior of that enzyme become available. In fact, this technique, Single-molecule Picometer Resolution Nanopore Tweezers (or SPRNT) [7], permits an order of magnitude improvement in spatiotemporal resolution over optical tweezers, a standard technology for single-molecule studies of nucleic acid processing enzymes. Using asymmetric salt profiles, SPRNT reaction conditions can also be optimized in the *cis* compartment, while maintaining high *trans* [K+] to obtain a high SNR. This study provides a guide for studying DNA motor enzymes at unprecedented resolution regardless of the enzymes preferred operating conditions.

0.4.2 RNAP transcription elongation tracked at high resolution

Introduction: SPRNT provides the first opportunity to watch RNAP movement at spatiotemporal resolution relevant to single RNAP steps at high [NTPs]. Single RNAP steps have previously been detected using single-molecule techniques, but only at low [NTPs] (1

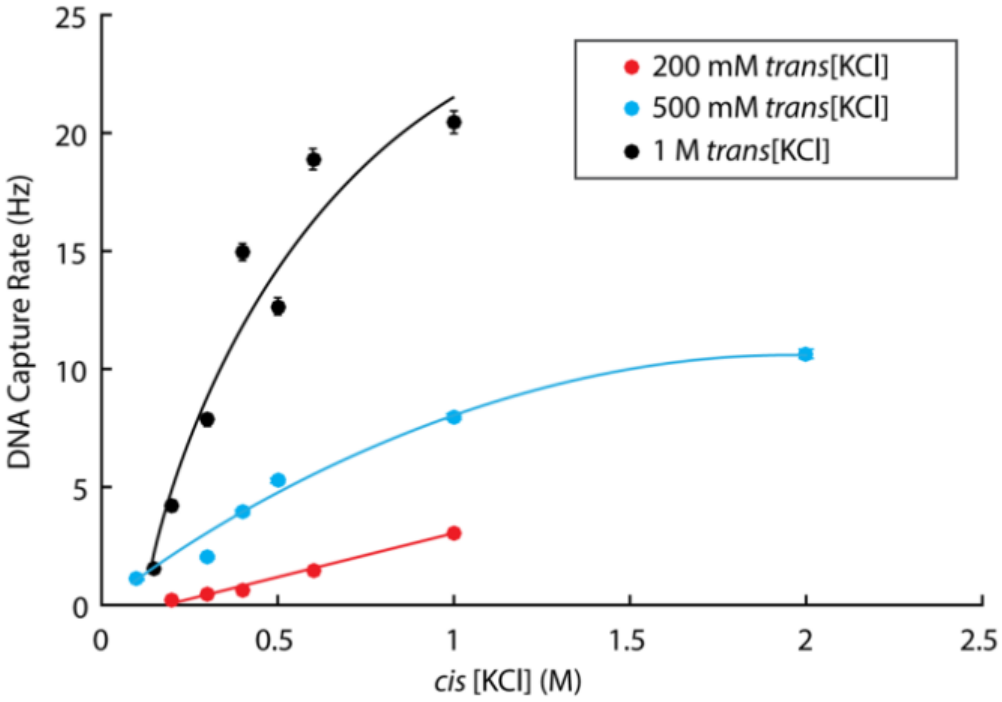


Figure 43: DNA capture rate, the number of DNA molecules threading through MspA per second, was measured using short hairpin DNA (500 nM) over a range of *cis* [KCl] at three *trans* [KCl] with an applied voltage of 180 mV. No phi29 DNAP enzyme was included in this set of experiments. Trend lines are to guide the eye. Errors are S.E.M.

DNA Sequences
All shown 5' > 3', X= abasic site, P= phosphate group, Z=cholesterol
Y=3 carbon spacer

seq1 template
PAAAAAAACCTTCCXCCCGTACCGTGCCGTACCGTTCCGTTCCGTACCGTAAATCAAGCA...
...ACTAATAAAAGCATTCTCATGCAGGTCGTAGCC

seq1 complement
TTTTATTAGTTGCTTGATTTACGGTACGGAACGGAAXXXXXXXY

seq 2 template
PAAAAAAACCTTCCXACACCGATTCTCCCGAGTCGGCCGAATCAAGCAACTAATAAAAGC...
...ATTCTCATGCAGGTCGTAGCC

seq2 complement
TTTTATTAGTTGCTTGATTCGGCCXXXXXXXY

Cholesterol primer (for both seq1 and seq2)
ZTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTXXXXXGGCTACGACCTGCAT..
...GAGAATGC

Hairpin Primer (for DNA Capture Rate experiments)
GCGTACGCCTACGGTTTTCCGTAGGCGTACGCGGCTACGACCTGCATGAGAATGC

Experimental Statistics

***trans* [KCl]**	***cis* [KCl]**	**DNA sequence**	**Number of pores**	**Number of events**
500 mM	300 mM	seq2	2	14
500 mM	225 mM	seq2	1	15
500 mM	100 mM	seq2	1	18
500 mM	50 mM	seq2	1	13
500 mM	0 mM	seq2	2	8
500 mM	400 mM	seq2	4	43
500 mM	150 mM	seq2	2	42
2 M	200 mM	seq1	2	24
1 M	200 mM	seq1	2	16
500 mM	200 mM	seq1	1	76
400 mM	200 mM	seq1	1	36
300 mM	200 mM	seq1	1	58
200 mM	200 mM	seq1	1	38
100 mM	200 mM	seq1	2	32
150 mM	200 mM	seq1	1	70
125 mM	200 mM	seq1	2	41

Figure 44: DNA sequences and experimental stats for asymmetric salt experiments

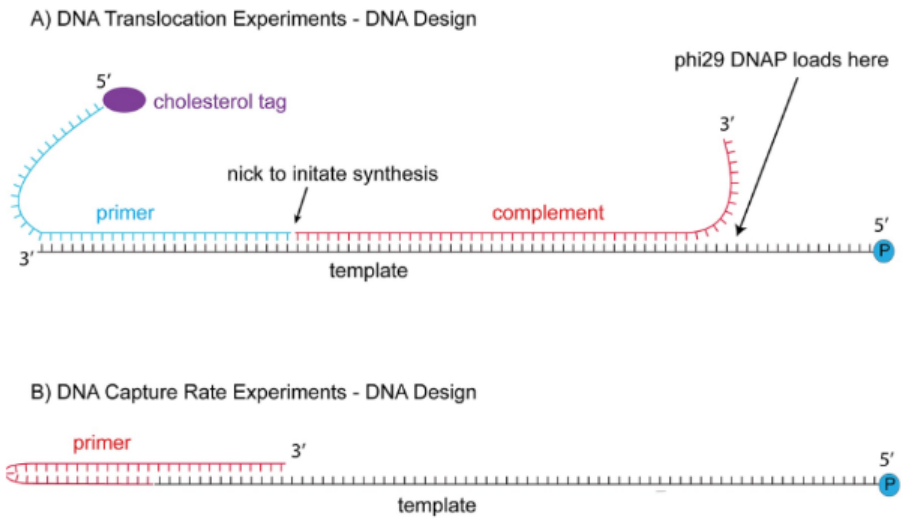


Figure 45: DNA design for asymmetric salt experiments

to 10 μM), where the rate of transcription is slowed to $\sim$1 base per second [4]. Not only are these measurements a far departure from physiological conditions, where the enzyme transcribes at $\sim$25 nt/s, but they only provide access to kinetic analysis where NTP binding is the rate limiting step in the reaction cycle. Observation of RNAP stepping over a full range of [NTPs] will be required to determine the full reaction pathway and relevant rate constants during transcription. In addition, previous kinetic modelling of RNAP suggests that the enzyme will oscillate between translocation states prior to NTP binding at low [NTP] ([4]). For this reason, even with slowing the transcription rate with low [NTPs], there may still be changes in position occurring at the rate of milliseconds. These oscillations will be unobserved with other single-molecule techniques.

Results: We acquired many reads of *E. coli* RNAP transcription over the same DNA scaffold using the methods outlined in *General Methods* and *Initial Results*. Figure 29 depicts three representative ion-current vs time traces acquired at 1000 μM NTPs, 21C, 180 mV (+ 38 pN). These ion-current vs time traces were converted into position vs time traces (Figure

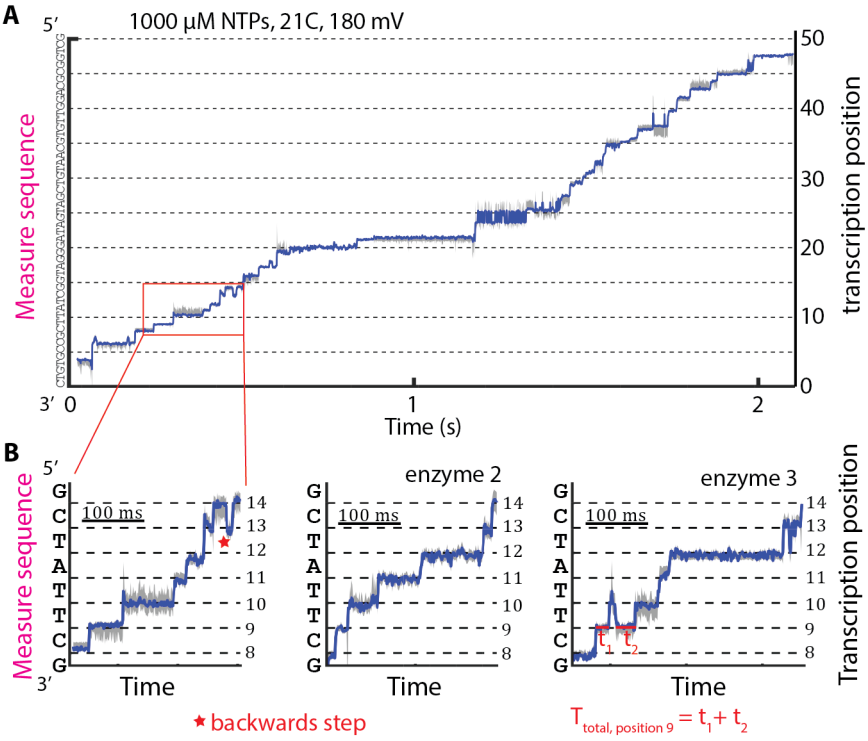


Figure 46: (A)SPRNT position trace for single enzyme transcribing 50 nts in ∼2 seconds. Grey depicts error in position calculation for every timepoint. (D) 7 nts of SPRNT position data from three individual RNAP enzymes transcribing the same sequence. Data plotted at 500 hz. Short duration enzyme steps and backwards transitions are detected (red star)

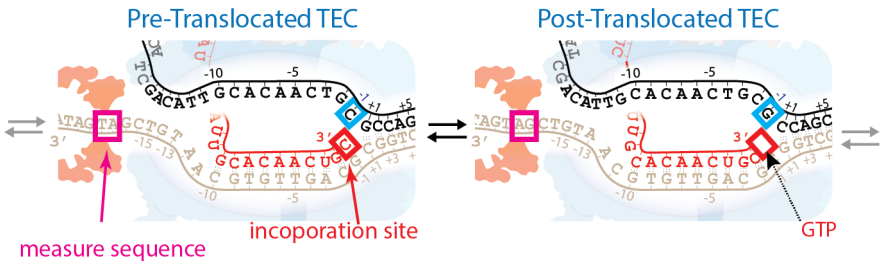


Figure 47: Schematic of Pre-translocated and Pos-translocated TECs. tDNA sequence in constriction region of MspA (measure sequence, pink box) determines ion current. Observable kinetic transitions (black arrows) change measure sequence. Other kinetic transitions (grey arrows: NTP binding, unbinding and condensation) do not shift measure sequence and are unobserved.

46A) using the consensus current pattern for the DNA sequence derived with PcrAX (see *Initial Results*). The position traces reveal discrete enzyme steps over $\sim$ 50 transcription positions with millisecond durations. As expected, the stepping rate is stochastic, with variable stepping rates across the sequence. By zooming in on a section of this trace (Figure 46 B), we observed discrete enzyme steps where the size of each step corresponded to a shift in the measured sequence by $\sim$ 1 nt. We also detected backward steps, where the enzyme moved backwards along the scaffold, shifting the tDNA measure sequence by $\sim$ 1 nt in the 3′ direction. At some positions, we also observed rapid oscillation between adjacent transcription positions before progression to the next step.

Discussion: These results present the first single-molecule measurements of single steps of *E. coli* RNAP with ms durations and demonstrate the capabilities of observing single transcription reaction cycles at biologically relevant NTP concentrations with this technique. Because the actual measurements in SPRNT were produced by the tDNA in the constriction of the nanopore, only parts of the transcription reaction cycle that correspond to physical motion of the tDNA relative to the enzyme can be observed. The transitions between Pre and Post (Figure 47) correspond to a shift forward by the enzyme, but NTP binding, condensation, transitioning the TEC into the next Pre state do not correspond to physical

motion of the TEC and cannot be observed directly. Therefore, in sections of these RNAP stepping plots with consecutive steps without backwards transitions, each step contains a Post state without a bound NTP, NTP binding, NTP condensation (Pre), and then physical motion of the TEC into the next Post state.

0.4.3 Determining the distance correction

Introduction: The tDNA sequence measured during SPRNT-RNAP was upstream of the TEC and outside of the region affecting kinetics (Fig 48). The position traces in Figure 46 therefore specifically tracked the measure sequence in the nanopore over time. By determining the distance between the measure sequence and the incorporation site, we can instead track the movement of DNA inside the enyme over time. We sought to calculate the distance correction, defined as the number of nts between the $5'$ end of the measure sequence and the -1 site (where the incorporation site sits in the TEC schematic). In previous nanopore sequencing experiments with DNA Polymerase (DNAP), a distance correction for phi29 DNAP was calculated by limiting the concentration of one of the four dNTPs and finding the positions where the durations increased [7]. The positions where durations increased corresponded to incorporation positions of the limited dNTP type.

Results: We slowed RNAP by lowering [UTP] to 10 μM while maintaining all three other [NTPs] at 1000 μM (n = 35 enzymes) (Figure 48A). We calculated the total time spent in each transcription position ($\langle\, T_{total} \,\rangle$) across 36 transcription positions. Gaps in the measurements in Figure 48A corresponded to positions with limited resolution between adjacent positions. At some positions, the $\langle\, T_{total} \,\rangle$ at low [UTP] was unchanged compared to $\langle\, T_{total} \,\rangle$ with 1000 μM of all four NTPs. These marked transcription positions not associated with U incorporation. However, at low [UTP], $\langle\, T_{total} \,\rangle$ increased at other transcription positions compared to $\langle\, T_{total} \,\rangle$ with 1000 μM of all four NTPs. We aligned the pattern of spikes in $\langle\, T_{total} \,\rangle$ with the positions of uracil in the RNA sequence, determining the transcription positions associated with UTP incorporation. This alignment produced a 17 nt distance

correction.

We corroborated this calculation by limiting the other three NTP types (GTP, CTP, ATP) sequentially and repeating the measurement independently, with each experiment confirming the 17 nt distance correction (Fig 48). Averaged over the whole sequence, the increase in $\langle\ T_{total}\ \rangle$ was greatest for U incorporation and smallest for C (Figure 48B). All of this data was recorded at a single applied voltage (180 mV).

Discussion: Using this distance correction calculation, we can now associate transcription positions with specific ribonucleotide incorporation along the scaffold, enabling high resolution investigation of RNAP at particular DNA sequences. For the SPRNT position *vs.* time traces, we can now align the identity of the nucleotides within the transcription bubble with each position measurement. Specifically, we track the -1 ntDNA over time (Figure 48C), as in the traces in Figure 48D. We analyze the dynamics of these type of traces in the next section.

The difference in the magnitude of the ratios of $\langle\ T_{total}\ \rangle$ between the different NTP types highlight the different rate constants for NTP binding and incorporation between the different types. In particular, the ratio of $\langle\ T_{total}\ \rangle$ for CTP was significantly lower than the other NTP types, suggesting that CTP has a larger NTP binding rate than the others. Additionally, even among the the different positions for a single NTP, there was considerable variation in the ratio of $\langle\ T_{total}\ \rangle$, confirming that the other nucleotides inside the enzyme play a significant role in incorporation kinetics.

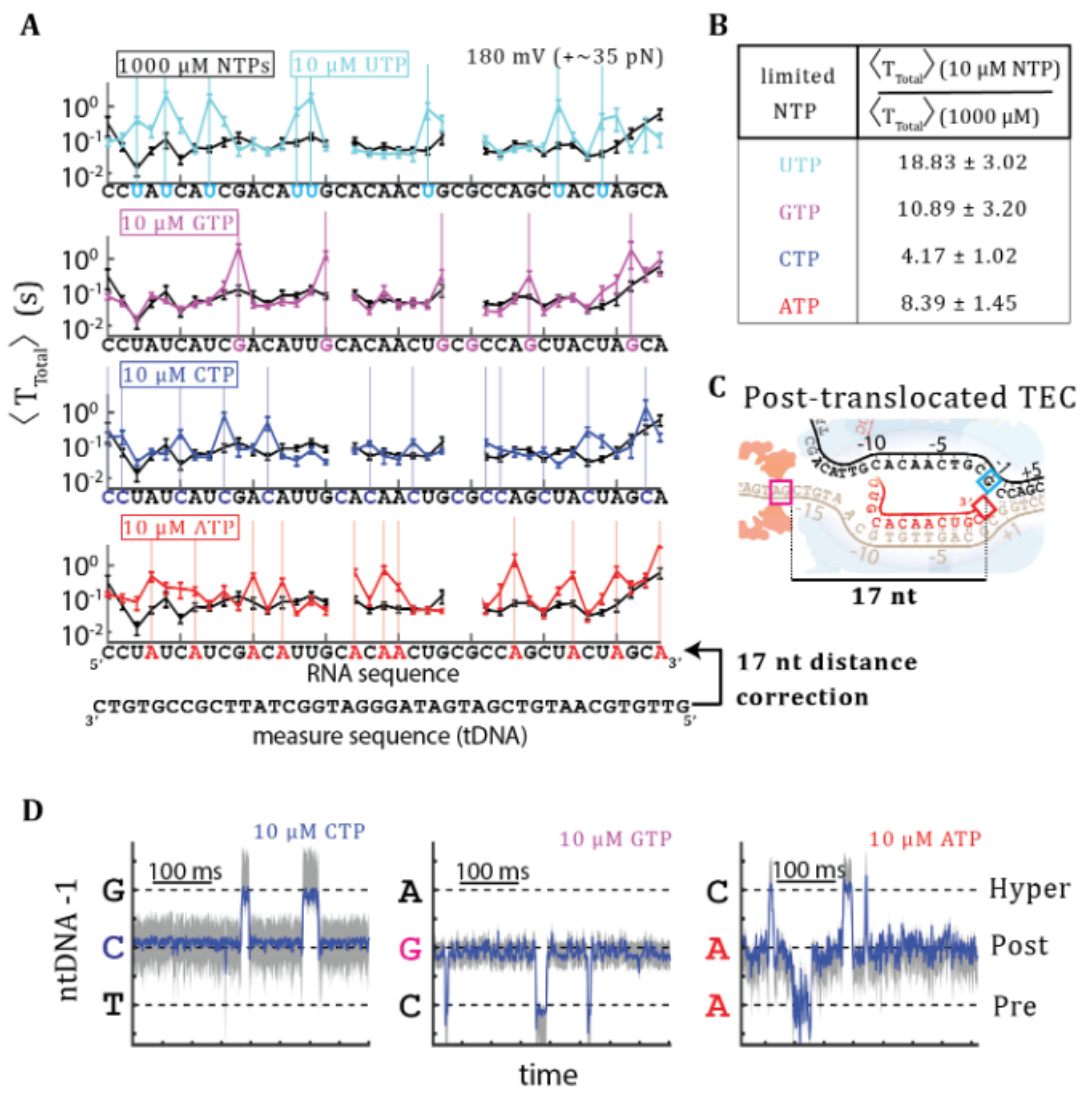


Figure 48: (A) ⟨ T_{total} calculated at each transcription position. RNA sequence aligned with each trace to maximize the average⟨ T_{total} (10 μM NTP) /⟨ T_{total} (1000 μM NTP) at positions with the limited NTP according to the alignment (17 nt distance correction). Gaps in plots at specific positions represent locations with either low resolution between consecutive measure sequences or elemental pause sequence locations (B) Table of average ⟨ T_{total} (10 μM NTP) / ⟨ T_{total} (1000 μM NTP) for each NTP type at positions with limited NTP type according to 17 nt distance correction. (C) Schematic of Post TEC with 17 nt distance correction labeled. ntDNA nt at -1 position identifies NTP type to be incorporated. (D) Example traces of TEC state transitions at three different transcription positions. Data plotted at 1 khz.

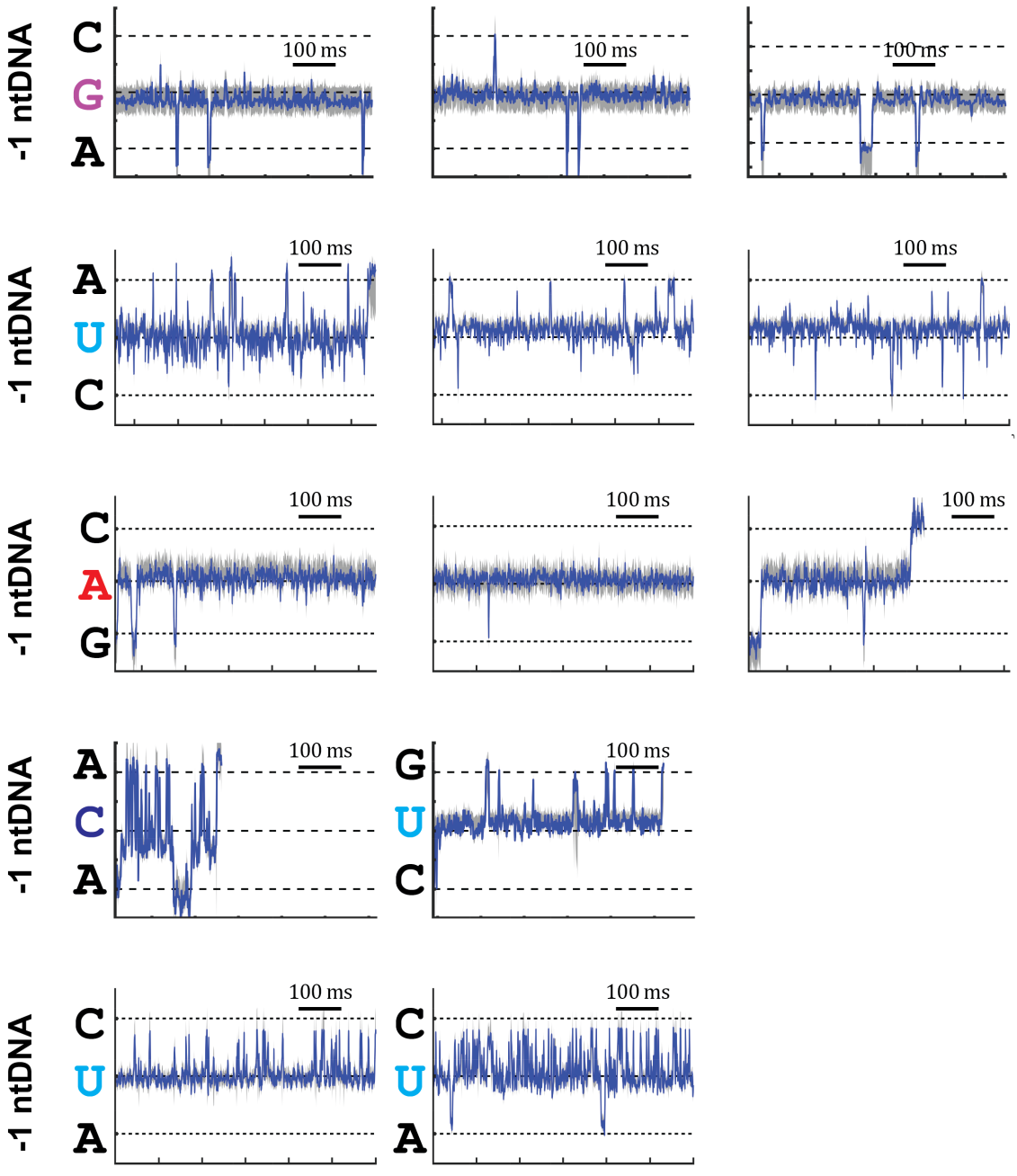
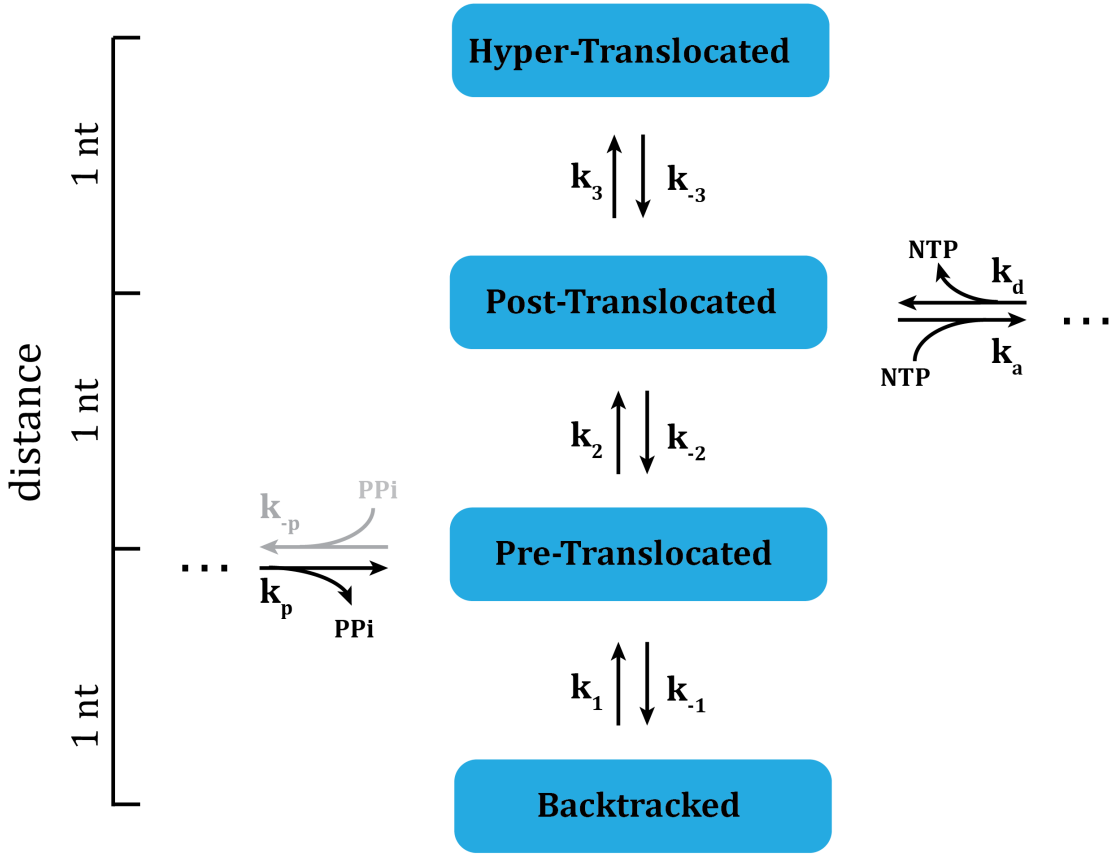

-1 ntDNA
C
G
A
100 ms
100 ms
100 ms
-1 ntDNA
A
U
C
100 ms
100 ms
100 ms
-1 ntDNA
C
A
G
100 ms
100 ms
100 ms
-1 ntDNA
A
C
A
100 ms
G
U
C
100 ms
-1 ntDNA
C
U
A
100 ms
C
U
A
100 ms

Figure 49: Position *vs* Time plots for many transcription positions at low [NTP]. Each plot was generated using a different *E. coli* RNAP enzyme, where the NTP type shown in color was reduced to 10 μM, while the other three [NTPs] were 1000 μM. The -1 ntDNA is aligned with the transcription positions according to the 17 nt distance correction between measure sequence and incorporation site. These plots show oscillations between multiple transcription positions at low [NTPs]. The behavior varies at different incorporation sites, with some states having a longer lived dwell time in the central position, and some positions favoring deviations forward and some backwards. The behavior was consistent across enzymes at the same transcription positions. All data was acquired at 180 mV, 21C. Data was plotted at 1 kHz.

0.4.4 Tracking RNAP state transitions during elongation

Introduction: The distance correction helped define the specific nucleotide incorporations associated with particular transcription positions. But for many of the transcription positions, the increase in $\langle\ T_{total}\ \rangle$ at low [NTP] spanned multiple adjacent positions for a single incorporation. This was clearly observable in the position traces at low [NTP], where the enzyme often oscillated (stepping forward and backward) between multiple transcription positions (Figure 48D) prior to progression forward completely (and Figure 49). We sought to quantify this behavior more thoroughly to assign these oscillations to specific TEC state transitions.

Transcription Reaction Schematic



$t_{Hyper} = 1/k_{-3}$

$t_{Post} = 1/(k_{-2}+k_3+k_a[NTP])$

$t_{Pre} = 1/(k_{-1}+k_2+k_{-p}[PPi]) = 1/(k_{-1}+k_2)$ as $[PPi]\sim 0$

$t_{Back} = 1/k_1$

Figure 50: Simplified schematic of a single transcription reaction cycle. Vertical transitions between different TEC translocation states correspond to physical motion of the TEC relative to the DNA and can be observed in SPRNT. Backtracked and Hyper are considered off-pathway states, as visits to these states require departure from the linear transcription pathway. The dwell time in each state is equal to the inverse of the sum of the rate constants out of that state. Because the concentration of Ppi is very low ($\sim$0), the rate backwards out of Pre is considered to be negligible and is plotted in grey. For all of the states (except Post) the dwell time is independent of the [NTP]. For the Post state, the dwell time should increase at low [NTP]. According to this model, visits to other TEC states out of Post should increase at low [NTP] as the rate of NTP binding decreases with [NTP]

According to kinetic models of RNAP transcription elongation (see *Background*), during a single transcription reaction cycle, the TEC can oscillate between various TEC states prior to NTP binding. The two primary TEC states are presumed to be Pretranslocated (Pre) and Postranslocated (Post), where NTP binding can only occur in Post. In addition, both Backtracked states, where the enzyme moves further backwards past Pre, or HyperTranslocated states (Hyper), where the TEC shifts forward past Post and the 3 end of the RNA moves past the active site have been detected in various transcription studies, but the relative rate constants governing the transitions between TEC states have only been loosely estimated.

Results: In addition to $\langle\ T_{total}\ \rangle$, we also calculated the average dwell time ($\langle\langle\ t\ \rangle$, defined as the average duration of every visit to each transcription position, Figure 52) and the probability of a backstep (P_{Back}, defined as the likelihood that a back step occurs at each transcription position in a given read, Figure 55) at every transcription position. We compared the ratios of all three kinetic parameters ($\langle\langle\ T_{total}\ \rangle$,$\langle\ t\ \rangle$,$P_{Back}$) between 1000 μM of all for NTP types and with limited NTPs (10 μM of each NTP sequentially with the other three unchanged (as in the previous section)) (Figures 51,54,55).

We defined certain transcription positions, that we call ii, according to the RNA sequence alignment with a 17 nt distance correction, such that positions ii equal transcription positions with the appropriate nucleotide type at [10 μM] (positions ii are marked with vertical lines in Figure 48A). Across all four nucleotide types, the average $\langle\ T_{total}$, 10 μM$\rangle$/$\langle\ T_{total}$, 1000 μM

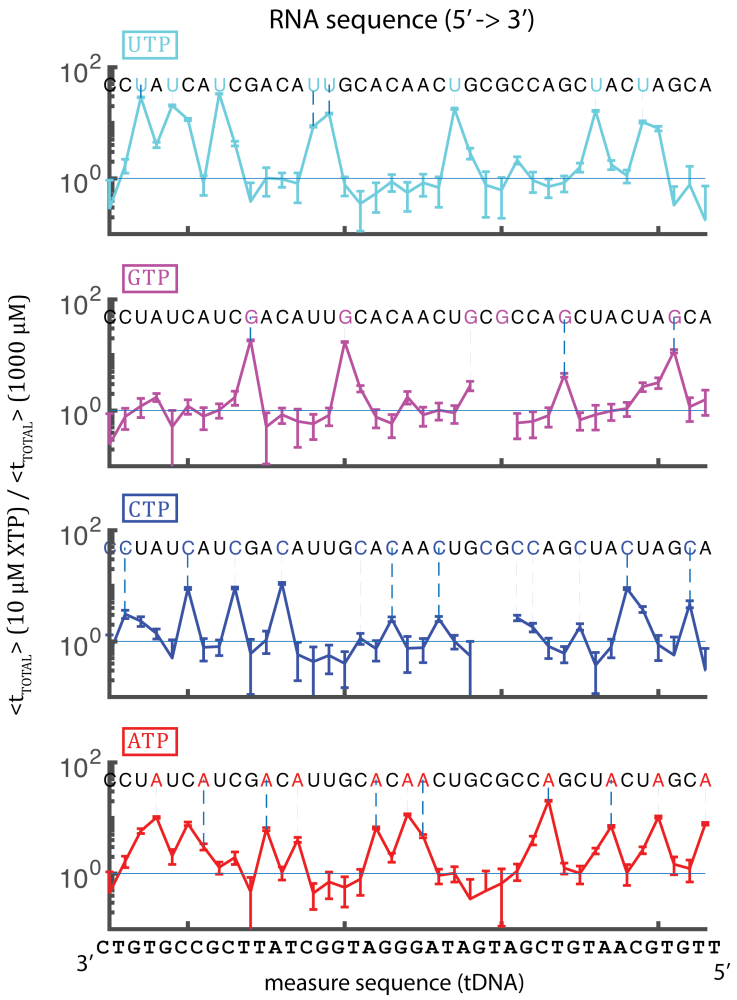


Figure 51: We calculated the ratio of $\langle\ T_{total}\ \rangle$ at limited [NTP] compared to saturating NTPs at every transcription position. We divided the $\langle\ T_{total}\ \rangle$ traces presented in Figure 48A. Data is plotted with a logarithmic y-axis. The RNA sequence is aligned to the plot by the distance correction (17 nt shift between measure sequence and RNA sequence) All of the positions where the NTP type is equal to the limited NTP type (vertical lines in plots) are classified as positions ii. We measured the greatest increase in the ratio of $\langle\ T_{total}\ \rangle$ at positions ii. However, the spikes extended to other positions adjacent, as the enzyme visits multiple states prior to NTP incorporation. This data is tabulated in the first row of Tables 56,61,60,59,58.

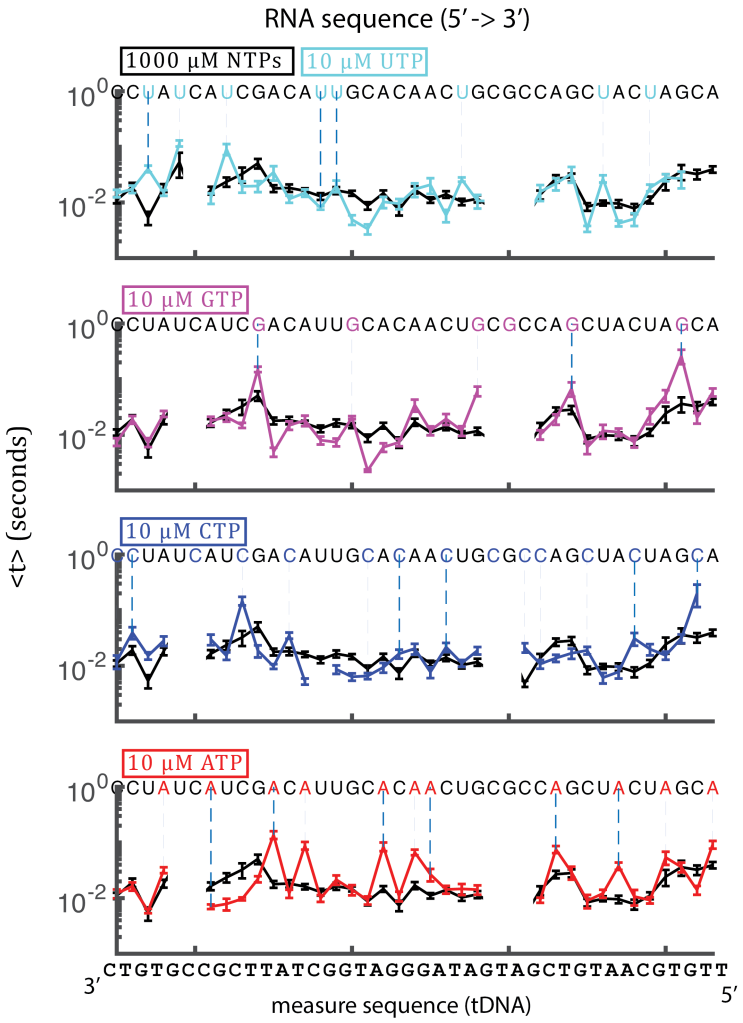


Figure 52: We calculated the average dwell time ($\langle$ t $\rangle$) at every transcription position using the same dataset as the calculations for $\langle$ T_{total} $\rangle$. The RNA sequence is aligned to the plot by the distance correction (17 nt shift between measure sequence and RNA sequence) All of the positions where the NTP type is equal to the limited NTP type (vertical lines in plots) are classified as positions ii.

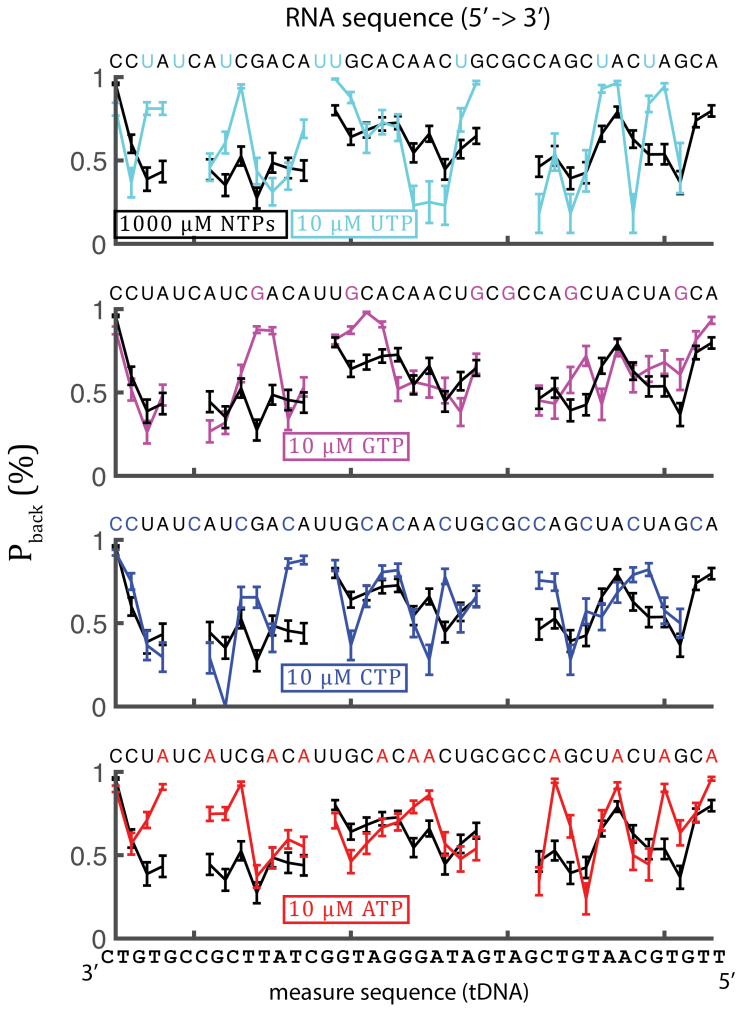


Figure 53: We calculated the backstep probability (P_{Back}) at every transcription position using the same dataset as the calculations for $\langle T_{total} \rangle$ and $\langle t \rangle$. The RNA sequence is aligned to the plot by the distance correction (17 nt shift between measure sequence and RNA sequence) All of the positions where the NTP type is equal to the limited NTP type (vertical lines in plots) are classified as positions ii.

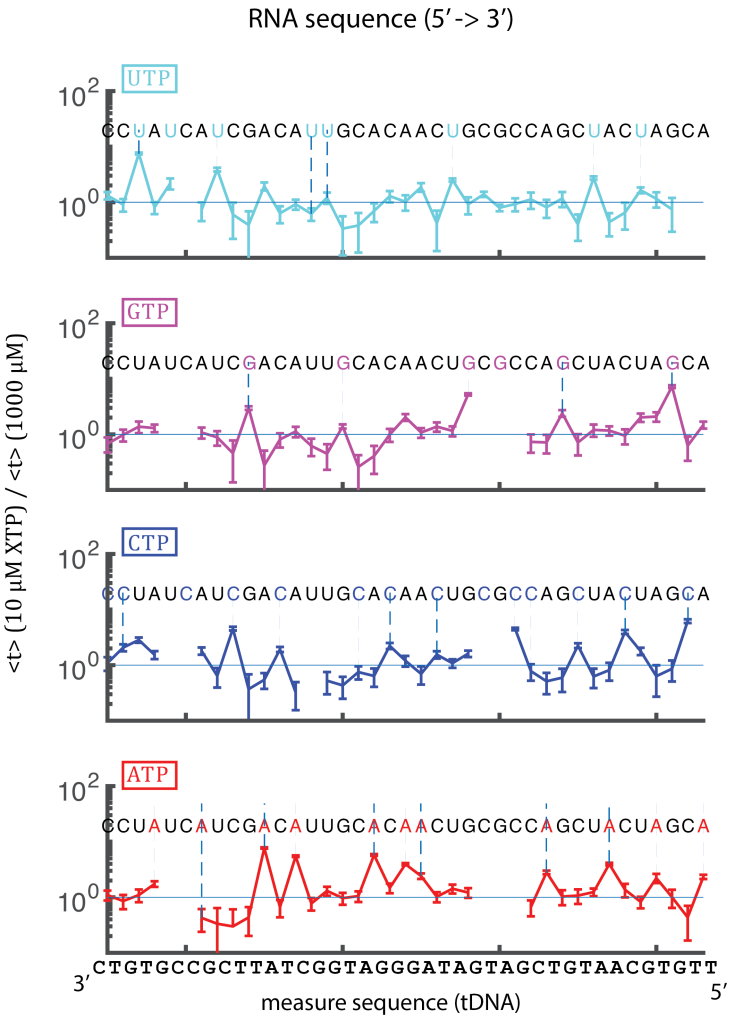


Figure 54: We calculated the ratio of $\langle$ t $\rangle$ at limited [NTP] compared to saturating NTPs at every transcription position. We divided the two $\langle$ t $\rangle$ traces presented in Figure 52. Data is plotted with a logarithmic y-axis. The RNA sequence is aligned to the plot by the distance correction (17 nt shift between measure sequence and RNA sequence) All of the positions where the NTP type is equal to the limited NTP type (vertical lines in plots) are classified as positions ii. This data is tabulated in the second row of Tables 56,61,60,59,58.

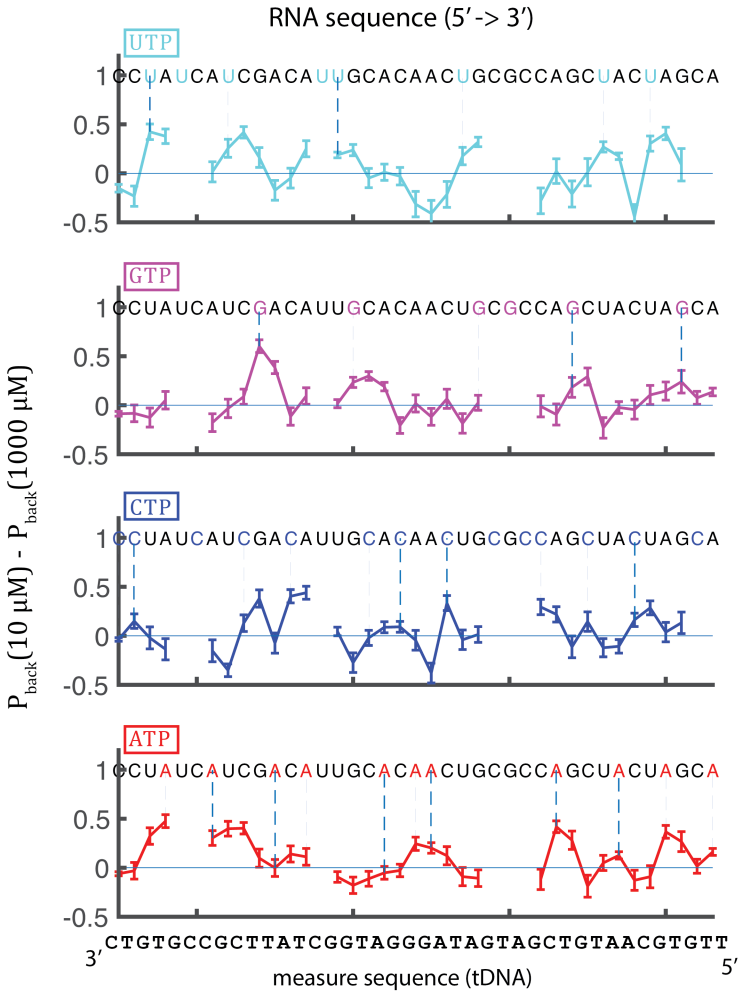


Figure 55: We calculated the difference in P_{Back} at limited [NTP] compared to saturating NTPs at every transcription position. We subtracted the two P_{Back} traces presented in Figure 53. Data is plotted with a logarithmic y-axis. The RNA sequence is aligned to the plot by the distance correction (17 nt shift between measure sequence and RNA sequence) All of the positions where the NTP type is equal to the limited NTP type (vertical lines in plots) are classified as positions ii. This data is tabulated in the third row of Tables 56,61,60,59,58.

$\rangle$ was largest at positions ii and was = 9.67 $\pm$ 1.30. The increase in $\langle$ T_{total} $\rangle$ often extended across multiple transcription positions. This was particularly apparent at 0 μM [GTP], where the increase in $\langle$ T_{total}, 0 μM$\rangle$/$\langle$ T_{total}, 1000 μM $\rangle$ spanned at least three transcription positions but was centered around positions ii. Along the same lines, we measured a small but significant increase in average average $\langle$ T_{total}, 10 μM$\rangle$/$\langle$ T_{total}, 1000 μM $\rangle$ at positions (ii +- 1) neighboring positions ii across all four nucleotide types. For example, at ii + 1 and ii - 1 we found average $\langle$ T_{total}, 10 μM$\rangle$/$\langle$ T_{total}, 1000 μM $\rangle$ was 1.43 $\pm$ 0.25 and 1.75 $\pm$ 0.32, respectively. At all other transcription positions (ii $\pm$ n, where n > 1), $\langle$ T_{total} $\rangle$ was unchanged (average $\langle$ T_{total}, 10 μM$\rangle$/$\langle$ T_{total}, 1000 μM $\rangle$ = 0.91 $\pm$ 0.07). This data is tabulated in Table 56.

We sought to assign these position measurements to TEC states. At positions ii, but not at ii + 1 or ii - 1, the average dwell time ($\langle$t$\rangle$) increased at 10 μM compared to 1000 μM for all four NTP types ($\langle$t ii, 10 μM$\rangle$/$\langle$t ii, 1000 μM $\rangle$ = 3.12 $\pm$ 0.33 at positions n) (Figure 54, Table 56). The dwell time in the Post state should depend on [NTP], as NTP binding and incorporation occurs in Post, whereas the dwell time in Pre (or off-pathway states (Backtracked, Hyper) , etc)) is only governed by the physical rate of transition back to Post, and is therefore [NTP] independent. This model is outlined in schematic Figure 50. In addition to the average dwell time, the dwell time histograms for positions ii were significantly altered (Figure 57) .

Thus, we conclude that in these SPRNT position measurements, during a single NTP incorporation cycle, position ii was the Post state, while visits backwards to ii - 1 were Pre and forwards to ii + 1 were Hyper. Across all the transcription positions measured at low [NTP], the average probability of entering Pre from Post prior to NTP incorporation ($P_{Back,ii-1,10uM}$ -$P_{Back,ii-1,1000uM}$ = 0.22 $\pm$ 0.03) increased by a similar amount to the average probability of entering Hyper from Post (Pback, ii+1, 10 μM - Pback, ii+1, 1000 μM = 0.19 $\pm$ 0.04), suggesting that the average rates of these two reactions (Post to Pre and Post to Hyper) are similar across all measured positions and at 180 mV applied voltage. Nevertheless, these relative probabilities varied significantly between transcription positions.

We also occasionally observed visits to other TEC states (Backtracked, Hyper +2), and the probabilities of these visits depended greatly on sequence position.

Discussion: These position traces present a general model for transcription elongation under a large (∼35 pN) assisting force. The relative reaction rates of the TEC state transitions out of Post prior to NTP incorporation vary significantly across transcription positions, with some positions heavily favoring transitions back to Pre over Hyper and vice versa. Because these measurements were made at a large assisting force, and previous single-molecule experiments have suggested transcription rate is force dependent, backward transitions (Post to Pre) may be reduced under our experimental conditions.

Methods:

$\langle \mathbf{T}_{total} \rangle$**:** For every read, we calculated the total time at each position by summing all of the dwell times of every visit to that position (T_{total}). We calculated the average T_{total} ($\langle T_{total} \rangle$) by taking the mean of T_{total} across all reads. We calculated the error in $\langle T_{total} \rangle$ using the standard error in the mean (s.d.m.) $= \sigma/\sqrt{n}$

$\langle \mathbf{t} \rangle$**:** For every visit to each transcription position across all reads, we calculated the dwell time and took the mean value of all dwell times for each position ($\langle t \rangle$). We calculated the error in $\langle t \rangle$ using the standard error in the mean (s.d.m.) $= \sigma/\sqrt{n}$

$\mathbf{P}_{Back}$**:** For each transcription position, we measured the total number of backwards steps (N_{Back}) across all reads (a single backstep from positon 4 to positon 3 counts as a backstep for position 4). We calculated the probability of a backstep for a given transcription position j using the following formula: $P_{Back,j} = N_{Back,j}/(N_{Back,j} + N_{reads})$ We calculated the error in $P_{Back,j}$ using the binomial distribution: $\sigma_{P_{Back,j}} = \sqrt{P_{Back,j} * (1 - P_{Back,j})/(N_{Back,j} + N_{reads})}$

	Position			
	ii-1	**ii**	**ii+1**	**other**
$\frac{\langle t_{TOTAL} \rangle (10\ \mu M)}{\langle t_{TOTAL} \rangle (1000\ \mu M)}$	1.43 ± 0.25	9.67 ± 1.30	1.75 ± 0.32	0.91 ± 0.07
$\frac{\langle t \rangle (10\ \mu M)}{\langle t \rangle (1000\ \mu M)}$	0.89 ± 0.09	3.12 ± 0.33	0.81 ± 0.08	1.11 ± 0.06
$P_{back}(10\ \mu M) - P_{back}(1000\ \mu M)$	-0.06 ± 0.04	0.22 ± 0.03	0.19 ± 0.04	0.03 ± 0.02

Figure 56: We tabulated the data from Figures 51 through 55. We took the averages of each kinetic parameter for all positions ii, ii+1, ii-1, and all other positions. The ratio of $\langle T_{total} \rangle$ (row 1) increased significantly at all three positions (ii, ii+1,ii-1) suggesting the enzyme spends significantly more time at all three positions during a single transcription reaction cycle at low [NTP]. For all other positions, the ratio was unchanged, as the [NTP] was still at 1000 uM (unchanged) for the other three NTP types which were incorporated at other transcription positions. The ratio of $\langle t \rangle$ (row 2) increased considerably at positions ii, but not at ii+1 and ii-1, suggesting that positions ii correspond to the Post translocated TEC, as only the dwell time in Post depends on [NTP] according to the model in Figure 50. Thus, positions ii-1 were Pre and ii+1 were Hyper. The difference in P_{Back} (row 3) increased at both positions ii and ii+1, where an increase at positions ii corresponded to more transitions from Post to Pre and an increase at positions ii+1 corresponded to more transitions from Hyper back to Post.

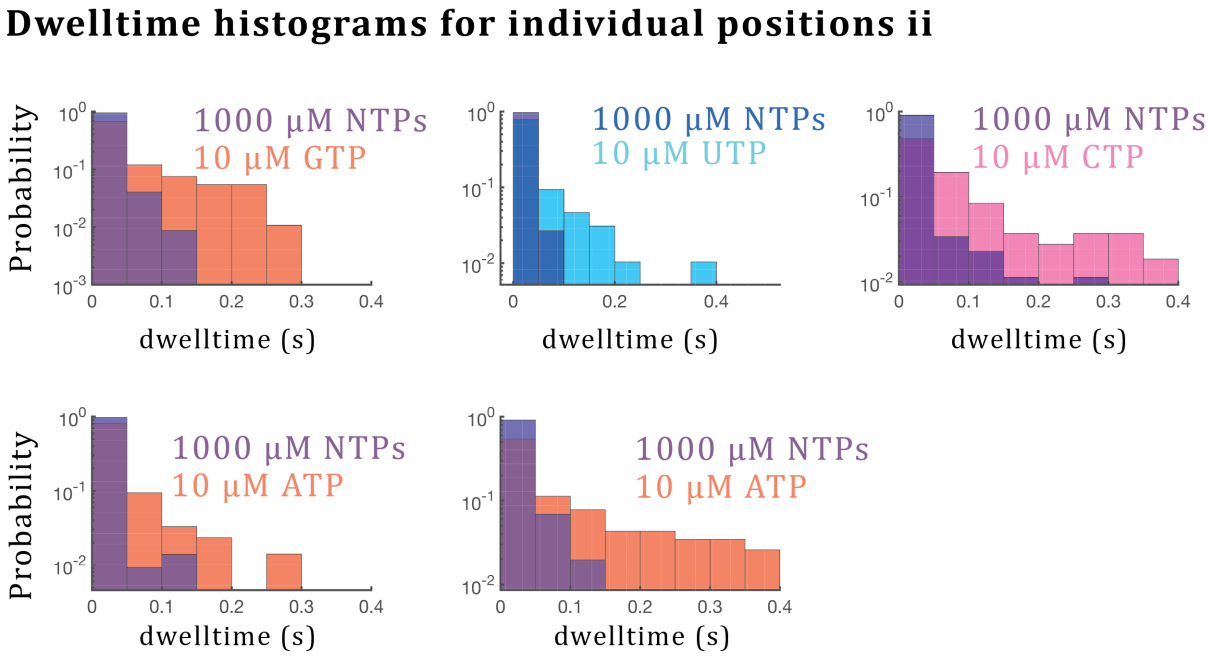


Figure 57: Using every measurement of the dwell time at a given transcription position, we created a dwell time histogram for five different transcription positions ii. The dwell time histograms for the positions ii at limited [NTP] exhibited a much longer tail and longer average dwell time compared to saturating [NTPs].

[UTP] =10 µM

	Position			
	ii-1	**ii**	**ii+1**	**other**
$\frac{\langle t_{TOTAL} \rangle (10\ \mu M)}{\langle t_{TOTAL} \rangle (1000\ \mu M)}$	1.13 ± 0.18	18.83 ± 3.02	3.60 ± 1.02	0.76 ± 0.11
$\frac{\langle t \rangle (10\ \mu M)}{\langle t \rangle (1000\ \mu M)}$	0.67 ± 0.09	2.79 ± 0.76	0.71 ± 0.13	1.03 ± 0.12
$P_{back}(10\ \mu M) - P_{back}(1000\ \mu M)$	-0.10 ± 0.10	0.27 ± 0.04	0.32 ± 0.04	-0.11 ± 0.04

Figure 58: We created the same table as Figure 56 except with solely data for limiting UTP experiments (top panel in Figures 51,54, 55). Across all of the UTP incorporation sites, the enzyme spent significantly more time in Hyper compared to Pre.

[GTP] =10 μM

	Position			
	ii-1	ii	ii+1	other
$\frac{<t_{TOTAL}> (10\ \mu M)}{<t_{TOTAL}> (1000\ \mu M)}$	1.49 ± 0.46	10.88 ± 3.19	1.21 ± 0.45	1.04 ± 0.12
$\frac{<t> (10\ \mu M)}{<t> (1000\ \mu M)}$	0.97 ± 0.30	3.85 ± 1.06	0.49 ± 0.11	1.11 ± 0.09
$P_{back}(10\ \mu M) - P_{back}(1000\ \mu M)$	-0.01 ± 0.06	0.26 ± 0.09	0.26 ± 0.07	-0.03 ± 0.03

Figure 59: We created the same table as Figure 56 except with solely data for limiting GTP experiments (second panel in Figures 51,54, 55)

[CTP] =10 μM

	Position			
	ii-1	ii	ii+1	other
$\frac{<t_{TOTAL}> (10\ \mu M)}{<t_{TOTAL}> (1000\ \mu M)}$	0.70 ± 0.07	4.17 ± 1.02	1.17 ± 0.33	0.94 ± 0.24
$\frac{<t> (10\ \mu M)}{<t> (1000\ \mu M)}$	0.77 ± 0.13	2.65 ± 0.50	1.12 ± 0.26	0.90 ± 0.32
$P_{back}(10\ \mu M) - P_{back}(1000\ \mu M)$	-0.14 ± 0.06	0.17 ± 0.04	0.10 ± 0.04	-0.02 ± 0.06

Figure 60: We created the same table as Figure 56 except with solely data for limiting CTP experiments (third panel in Figures 51,54, 55)

[ATP] =10 μM

	Position			
	ii-1	ii	ii+1	other
$\frac{<t_{TOTAL}>\ (10\ \mu M)}{<t_{TOTAL}>\ (1000\ \mu M)}$	2.50 ± 0.73	8.39 ± 1.45	1.29 ± 0.13	0.91 ± 0.16
$\frac{<t>\ (10\ \mu M)}{<t>\ (1000\ \mu M)}$	0.82 ± 0.12	3.58 ± 0.64	0.99 ± 0.15	1.25 ± 0.15
$P_{back}(10\ \mu M) - P_{back}(1000\ \mu M)$	0.02 ± 0.06	0.21 ± 0.05	0.15 ± 0.07	-0.04 ± 0.06

Figure 61: We created the same table as Figure 56 except with solely data for limiting ATP experiments (fourth panel in Figures 51,54, 55) Across all of the ATP incorporation sites, the enzyme spent significantly more time in Pre compared to Hyper.

0.4.4.1 Alternative stepping behaviour

While the commonly observed TEC states at low [NTP] were primarily Pre, Post and Hyper, we also more rarely observed enzyme visits to other TEC states in our measurements. Figures 62 and 63 depict examples of these alternate state visits. In Figure 62, at the first ATP incorporation site at 10 μM [ATP], we observed consistent oscillations into a state 2 nts in distance forward from Post. We define this other state Hyper (+2), where the enzyme shifts forward another full state past Hyper (+1) leaving two of the RNA:DNA hybrid contacts unfilled. At this particular position, we often observed oscillations to Hyper (+2) directly from Post, skipping over the Hyper (+1) state. This behavior was also clearly observed in the P_{back} measurements at low [ATP]. At this particular position, there was a large increase in P_{back} at position ii+2, corresponding to increased visits to a Hyper +2 state. We also occasionally observed Backtracked states (not shown), where the TEC shifted backward past the Pre state one or more positions.

These observations further highlight the wide diversity of behavior across incorporation

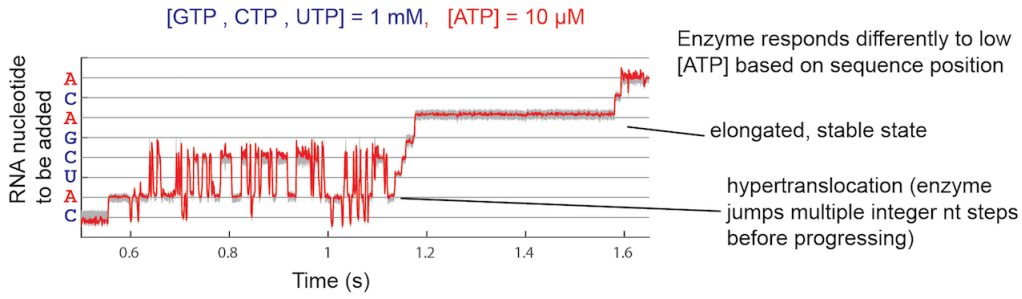


Figure 62: At limited [NTPs], we observed diverse behavior at different incorporation positions. While across the whole scaffold we mostly measured oscillations between Pre, Post and Hyper (section 0.4.4), at particular positions we also observed consistent oscillations from Post directly into further off pathway TEC states (Backtracked and Hyper +n). In the position plot shown here, at limited [ATP], the enzyme oscillated between Pre, Post and Hyper (+2) at the first ATP incorporation site. In contrast, the enzyme only visited the Post state for an extended duration at the next ATP site. These behaviors at these two positions were consistent across RNAP enzymes.

positions and different DNA sequence scaffolds. While some positions may have a very stable Post state, where the probability of oscillations to other states, even at low [NTP] are low (see second A incorporation position in Figure 62), other positions may oscillate between 2, 3, or even 4 separate TEC states. From a practical standpoint, this can make interpreting SPRNT position traces at low [NTP] difficult, particularly around the transitions between incorporations, as the Hyper state for one incorporation position is the same measured position as the Post state for the next.

Additionally, we also detected visits to Half-translocated states (red stars in Figure 63) in some position traces, where a full integer step was followed by two half integer steps, leaving one of the states halfway in between two fully translocated states. There is some evidence for Half-translocated intermediates in between Pre and Post in structural data of RNAP [82] [81]. However, all of the *e. Coli* RNAP structural half states were detected at elemental pause sequences, and it is an open question whether these states are visited during the normal transcription elongation pathway. While we did see occasional visits to Half-translocated intermediates (Figure 63) outside of pausing, the existence of these states was rare and only occurred transiently at particular transcription positions. Quantification

of this behavior was prohibited due to the noise in our position measurements at many positions and the brief durations ($\sim$ 1 ms) of many of the detected half states. Therefore, it could be the case that the Half-translocated state is an intermediate along the transcription reaction pathway, but is only visited transiently at certain sequences. Deeper analysis of this behavior will be required to make definitive claims about Half state dynamics outside of pausing. Half-translocated states during pausing are discussed in detail later in this document.

0.4.5 Effect of [NTP] on transcription rates (a second pause site).

By tracking many RNAP enzymes (n = 37) transcribing the same sequence (Figure 65), we measured kinetic rates for each transcription position in addition to the average kinetic parameters across all positions at saturating [NTPs]. As discussed previously, other than edge effects due to return from arrest and reaching the end of the sequence, $\langle \mathrm{T}_{total} \rangle$ at 1000 μM was relatively consistent from positions 1 to positons 23, with the exception of positions 24 and 25, where RNAP encounters the yrbL pause sequence (see section *Elemental pausing* for a depper discussion of this behavior). Outside of the pause location, $\langle \mathrm{T}_{total} \rangle$ at 1000 μM NTPs (180 mV) was 61 $\pm$ 24 ms averaged over all measured transcription positions with a range from 13.6 $\pm$ 5.6 (s.d.m) ms to 128 $\pm$ 28 ms (s.d.m).

When we lowered all [NTPs] to 100 μM (n=28 reads), $\langle \mathrm{T}_{total} \rangle$ increased at some transcription positions, but was unchanged at others, suggesting that the saturation point for NTP binding is below 100 μM for certain transcription positions. Interestingly, there was a second sequence location (in addition to the yrbL sequence), where $\langle \mathrm{T}_{total} \rangle$ increased significantly at 100 μM. $\langle \mathrm{T}_{total} \rangle$ at these positions (8 and 9) was comparable to $\langle \mathrm{T}_{total} \rangle$ at the yrbL position, suggesting that there is a second pause site in this sequence below 100 μM [NTP]. This second sequence does not match the consensus pause sequence. Instead, it corresponds to the Post state during the incorporation of two consecutive Uracil bases. In transcription termination, a long track of U incorporation is hypothesized to induce enzyme pausing. Perhaps the second pause site in the sequence measured here at low [NTP] functions

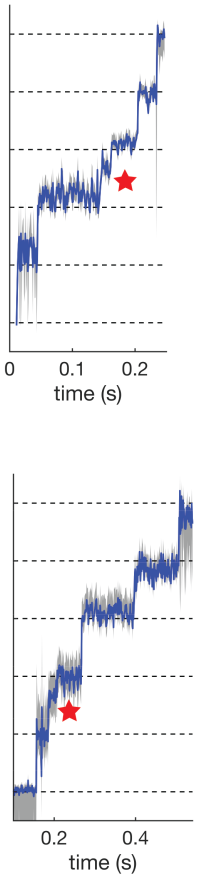


Figure 63: We also observed half-translocated steps in some of the SPRNT position traces. The red stars denote positions of Half steps, where the average position of the discrete step lies in between two fully translocated steps. It was difficult to quantify the probability of observing this behavior at each position, as these half-steps were rare and very brief in duration. This suggests that a Half translocated state was not a requisite transition along the reaction pathway but may exist transiently at certain positions. See section *Elemental Pausing* for a deeper investigation into Half TEC states.

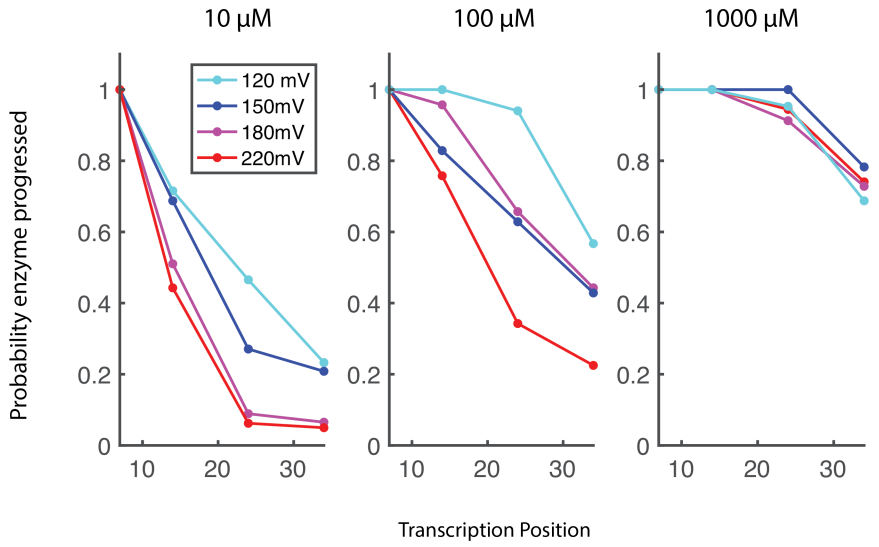


Figure 64: Many of the SPRNT - RNAP reads ended before RNAP had reached the end of the nucleic acid scaffold. At each voltage and each [NTP], we calculated the probability that the enzyme reached four separate transcription positions. The probability the enzyme progressed (a metric for RNAP processivity), dropped off quickly at 10 μM [NTPs] (Panel 1), particularly at higher voltages. In contrast, at 1000 μM NTPs (Panel 3), $\sim$75% of the RNAPs reached transcription position 37, regardless of applied voltage. From these results, we conclude that RNAP processivity with SPRNT was dependent on both [NTP] and voltage.

under a similar mechanism as pausing in transcription termination.

0.4.6 Effect of Varying Force on RNAP Transcription Kinetics

Introduction: After determining the distance correction and measuring the average transcription kinetics at a single voltage (i. e. applied force, 180 mV, +35 pN), we investigated the effect of applied force on transcription elongation kinetics with SPRNT. In single molecule optical tweezers experiments, transcription rate of *E. coli* RNAP core enzyme was dependent on the magnitude of the applied force in the experiment [4]. These force *vs.* rate relationships were used to differentiate between various kinetic models of transcription elongation (see *Background*).

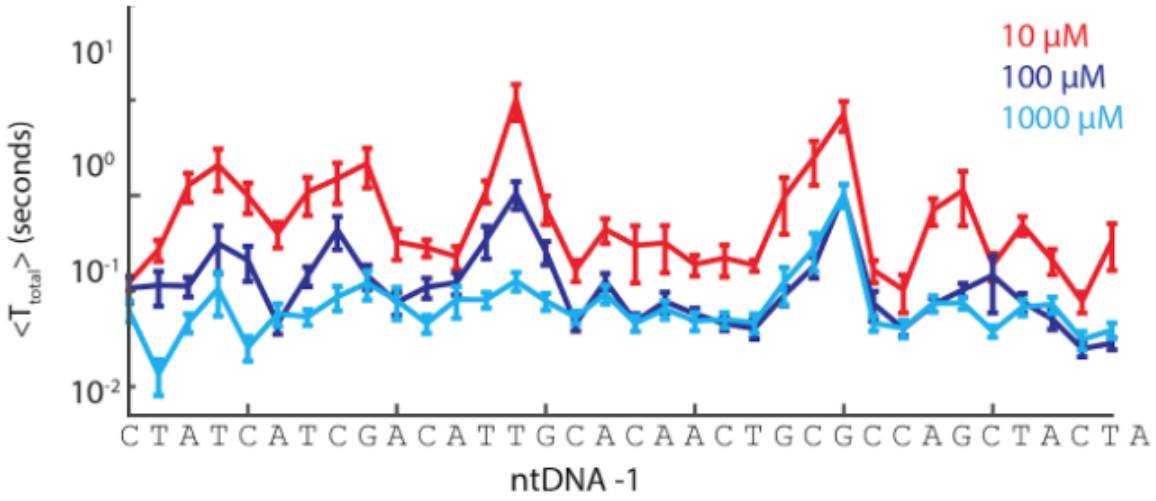


Figure 65: T_{Total} was calculated at each transcription position across [NTPs] (10, 100, and 1000 μM) at a single applied voltage (180 mV). The ntDNA -1 sequence was aligned to the transcription positions according to the 17nt distance correction defined in section 0.4.3

0.4.6.1 Transcription Rate vs. Force

Results: Due to the factors addressed in the section: ′implications of varying force′, at many transcription positions it was difficult to maintain single - step resolution over the full range of applied voltage (120 mV - 220 mV). For this reason, we first tracked the change in average transcription rate with force over multiple transcription reaction cycles. Using the same DNA scaffold sequence, we gathered many SPRNT-RNAP reads over a range of applied voltage (120 mV, 150 mV, 180 mV, 220 mV) and over a range of [NTPs] (10 μM, 100 μM, 1000 μM). We divided each SPRNT - RNAP read into specific sections of varying length (transcription sections) that were clearly identifiable over the whole range of applied voltage (Figure 66). For all of the transcription sections, with 10 μM NTPs, average transcription rate increased significantly over the range of applied voltage. Both the range of average transcription rates and the relative change in rate between the lowest (120 mV) and highest (220 mV) measured voltages varied significantly between the different transcription sections. We also measured an increase of transcription rate with force at the higher [NTPs]. At

1000 μM NTPs, the relationship between transcription rate and force began to saturate (the transcription rate was stable as the voltage was increased).

Note on various ways to calculate transcription rate: For each read we calculated the time spent in each transcription section. Various methods can be used to calculate the average transcription rate (Figure 68). In method 1, we calculated the average time spent in each transcription section. We then divided this average by the number of transcription positions in each section, yielding an average transcription rate. In method 2, we calculated a transcription rate for each enzyme for each section by dividing the time spent by the number of transcription positions in each section. We calculated an average transcription rate by finding the mean of the transcription rates for each section. Method 2 was more directly comparable to the method used to generate the force vs. velocity curves in optical tweezers experiments (see section *Comparing SPRNT to Optical tweezers*).

These two methods yielded different results, with method 2 always producing a faster transcription rate compared to method 1 Figure 68. The tail of the distributions in the two methods shift the mean value in opposite directions. We believe method 1 was a more accurate way to calculate the average transcription rate, as the actual measured parameter in SPRNT was time spent at each position, not transcription rate. Nevertheless, the average transcription rate calculated using method 1 can be affected significantly by individual measurements where the enzyme stalls at individual transcription positions. For this reason, a large sample of measurements was required to accurately determine the mean transcription rate. Therefore, we also calculated the median transcription rate (which gave the same result with either method 1 or method 2) to most accurately measure the relative change in transcription rate between voltages (Figures 69 and 70).

0.4.6.2 Transcription Rate vs. Force (at individual positions)

Introduction: The results from section 0.4.6.1 confirmed that overall transcription rate increases with applied force across all of the different sections of sequence. We next investigated the effect of applied force on individual incorporation positions. We identified

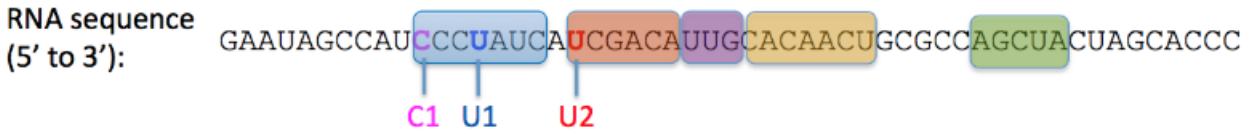


Figure 66: We divided the incorporation positions into sections that were easy to identify at each applied voltage based on the underlying ion-current patterns. Each section is marked by a different color. We also marked three individual incorporation positions (C1,U1, and U2) where resolution between the adjacent positions was maintained over the full range of applied voltage (120 mV to 220 mV).

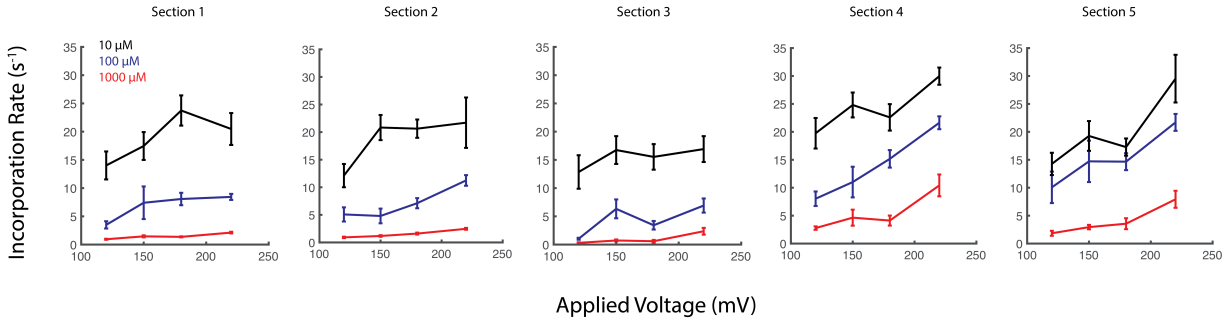


Figure 67: For each of the sections highlighted in Figure 66, we calculated the average incorporation rate across the section over a range of [NTPs] (10, 100, 1000 μM) and applied voltages (120, 150, 180, 220 mV).

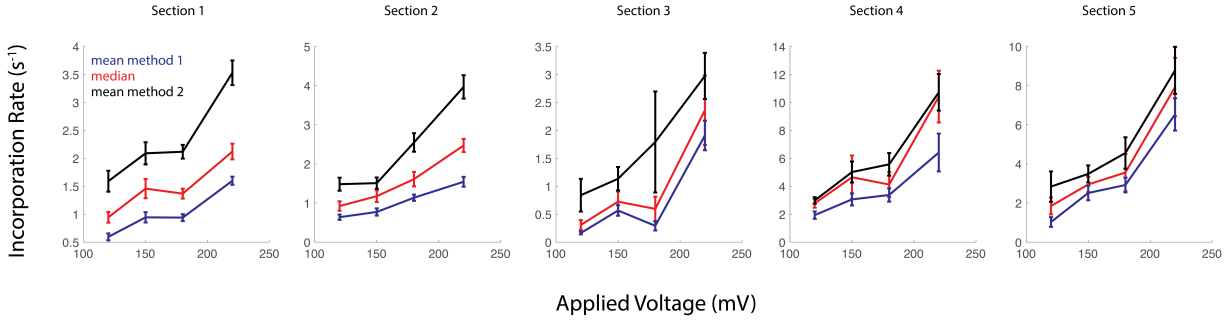


Figure 68: We compared different methods to calculate the incorporation rates (mean method 1, median, and mean method 2; the details of these calculations are outlined in the main text.). At 10 μM [NTPs], mean method 2 always produced a larger measurement than mean method 1. Even though mean method 1 was the most accurate representation of our data, mean method 2 was used to compare SPRNT data to optical tweezers data in Figure 72 because the optical tweezers experiments used mean method 2.

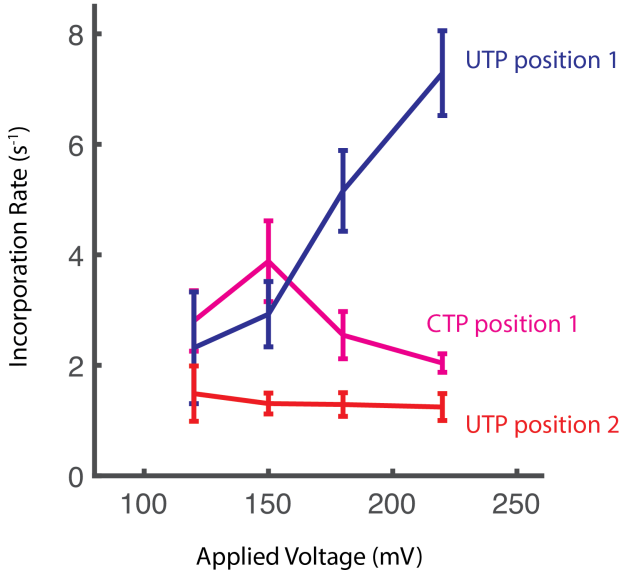


Figure 69: Incorporation rate (median method from Figure 68) was calculated across the range of applied voltages and at 10 μM NTPs for each of the three individual incorporation positions (C1, U1 and U2). The number of enzymes used for each measurement was outlined in Figure 71.

incorporation positions where differentiation between the transcription states (Pre, Post, and Hyper) was possible across the entire range of applied voltage (120 mV to 220 mV). For many of the incorporation positions, particularly at the lower voltages, clear identification of all three states was prohibited (see section 0.3.4 for rationale on this phenomena). In addition, because the enzyme processivity was low at 10 μM [NTPs], particularly at high voltages (Figure 64), we chose incorporation positions early along the sequence to maintain experimental throughput.

Results: Using this criterion (resolution between Pre, Post, and Hyper at all voltages AND early in sequence to maintain throughput at low [NTP]), we identified three incorporation positions (C1, U1 and U2 in Figure 66) for this analysis. For every SPRNT read, at all three positions and at 10 μM, we calculated the incorporation time, defined as the total time spent in Pre, Post and Hyper at that incorporation position. At each voltage, we calculated the

median incorporation rate, defined as the inverse of the median incorporation time (Figure 69). Interestingly, the incorporation rate had a very different response to changing the applied voltage across the three positions. For position U1, the incorporation rate increased across the whole range of voltage. In contrast, for both C1 and U2, the incorporation rate did not increase with applied voltage. For C1, the incorporation rate actually decreased slightly at the higher voltages, suggesting that an assisting force can slow down transcription at certain positions. For U2, the incorporation rate was relatively unchanged across voltages.

While each position had a unique response to changing the applied voltage, the average of the incorporation rate at each voltage across these three incorporation positions still increased with applied voltage (Figure 70). When compared to the rate *vs.* voltage curves for all of the sections of sequence at 10 μM, the average of the three positions was within the range of absolute values and slopes. The number of SPRNT reads used for every incorporation rate calculation are outlined in Table 71.

Discussion: Because the change in average transcription rate across all three positions was consistent with the change in rate across the sections of sequence (Figure 70), we believe these three incorporation positions were reasonably representative of the various kinetic situations for RNAP during transcription. Over particular ranges of assisting force, RNAP incorporation rate can either increase, decrease, or remain relatively unchanged depending on the RNAP position along the scaffold. On the surface, the concept of transcription rate decreasing as the assisting force increases (position C1, Figure 69) seems paradoxical. However, the existence of an off-pathway state which has an entrance probability that increases with assisting force can explain this behavior. The Hyper state clearly fits this criterion, as entrance into Hyper from Post should increase with assisting force, decreasing the amount of time spent in Post, and lowering the probability of NTP binding. Overall, the diverse coupling between applied voltage and transcription rate at the various incorporation positions was the result of the diversity of underlying rate constants at each position between Pre, Post and Hyper. In the next section, we will describe experiments aimed at determining

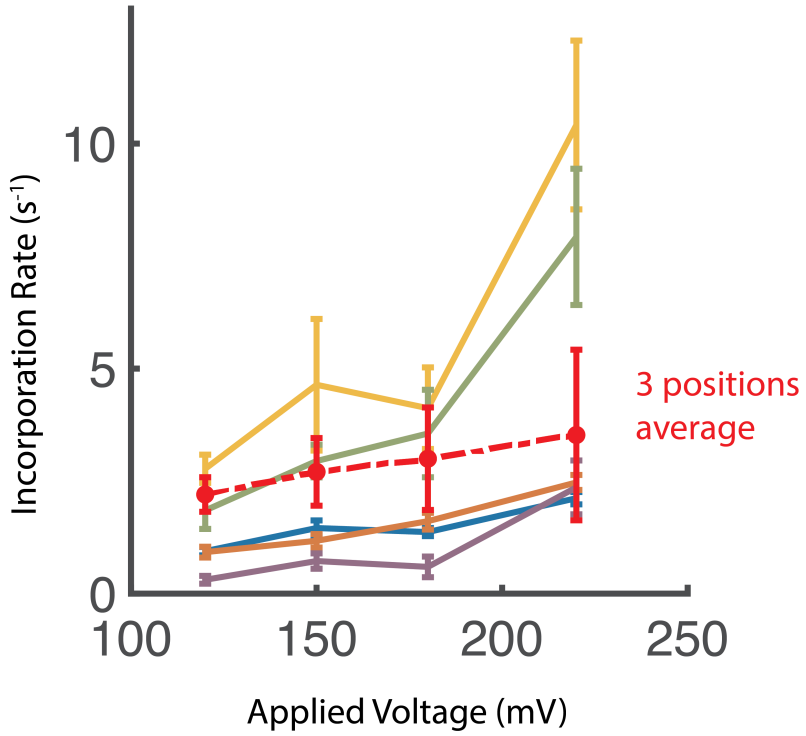


Figure 70: Incorporation rate *vs.* voltage for each section of sequence in Figure 67 at 10 μM NTPs. The colors of the traces correspond to the sections specified in Figures 66 and show median incorporation rate for each section and each voltage. Plotted in red was the average of the three position measurements from figure Figure 69.

Voltage (mV)	C1: n (durations)	U1: n (durations)	U2: n (durations)
120	115	67	76
150	106	91	81
180	260	197	149
220	286	198	83

Figure 71: Experimental counts for the data used to measure the average incorporation rate for the individual incorporation positions (C1,U1 and U2). Each count represents an individual measurement of the rate of transcription for each section from a single RNAP enzyme.

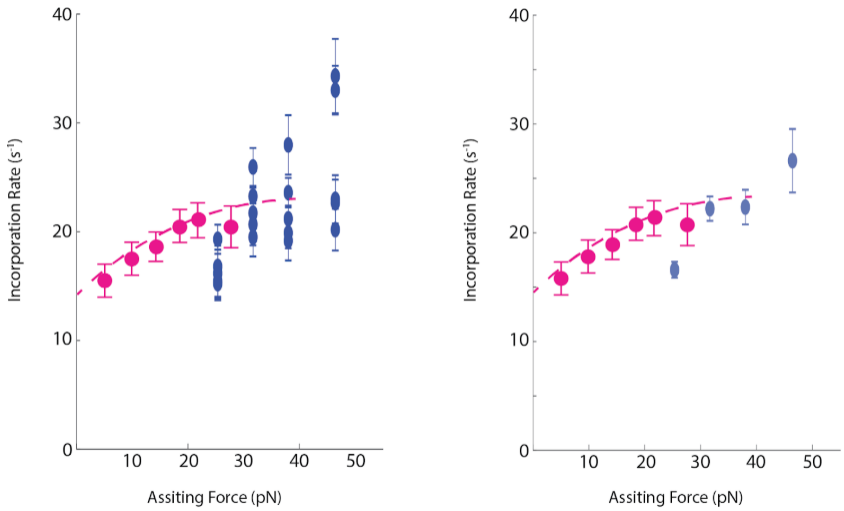


Figure 72: We used the SPRNT - force estimates to compare the incorporation rate vs. force curves derived with SPRNT to those obtained with optical tweezers. The data was acquired with 1000 μM NTPs and incorporation rate was calculated using mean method 2. Panel 1 shows the SPRNT measurements for each section of sequence independently. Panel 2 depicts the average of the SPRNT data in Panel 1. The optical tweezers data was taken from Figure 4 [4].

these rate constants.

0.4.6.3 Comparing SPRNT measurements to optical tweezers

These measurements provide the first direct comparison between optical tweezers data and SPRNT data, as the full velocity *vs.* force curves at various [NTPs] have been published for *E. coli* RNAP core enzyme with optical tweezers [4]. In this optical tweezers experiment, Abbondanzieri *et. al* measured the transcription velocity across a range of forces with a ratio of [NTP] (termed $[NTP]_{eq}$), where each NTP type was at a different concentration (2:5:10:10 μM [CTP]:[ATP]:[GTP]:[UTP]). For this reason, our 10 μM [NTP] or 100 μM results were difficult to directly compare. Our 1000 μM results were the most similar experimentally as saturating conditions (250X the $[NTP]_{eq}$ ratio) were measured in the optical tweezers study.

We converted the rate *vs.* voltage curves derived with SPRNT at 1000 μM [NTP] to rate

vs. force, using the force estimates for SPRNT (1mV $\sim$ 0.211 pN, section 0.3.4). We plotted this data with the 250X $[NTP]_{eq}$ data in Figure 5A of [4] (Figure 72), including the fit line from the previous study. Our measurements of incorporation rate at 1000 μM qualitatively match the measurements with optical tweezers Figure 72). The left panel in figure 72 depicts the measured SPRNT values for each section of sequence, while the right panel depicts the average across the various sections. The main divergence occurred at 120 mV ($\sim$25 pN), where the average rate across sections measured with SPRNT was lower than that of optical tweezers. It could be the case that the differences between the DNA scaffolds used to measure rate in the two experiments produces a different shaped curve. Alternatively, the SPRNT force estimate could be an overestimation, as the rate *vs.* force curves may match better qualitatively by shifting the SPRNT data two the left along the x-axis. Nevertheless, these results suggest that the SPRNT force estimate is a reasonable value to assign the general range of forces applied in these experiments.

0.4.6.4 Determining underlying rates during transcription

Introduction: The results of the limiting NTP experiments at a single applied voltage (section 0.4.4) revealed that the TEC oscillates between Pre, Post, and Hyper prior to NTP binding. The rate constants governing these transitions were sequence dependent, as the transition probabilities varied widely between different transcription positions along the length of the scaffold. While these results present new information about the transcription elongation cycle, they were acquired at a high assisting force (180 mV $\sim$ 38 pN), which pushes the TEC forward throughout the course of the reaction. Because the transitions between TEC states were the result of physical motion of the RNAP enzyme along the DNA, the magnitude of the applied force should alter the rate constants governing the transitions between TEC states.

The results presented in section 0.4.6.1, demonstrated that, averaged over sections of transcription, the transcription rate increased with applied voltage between 120 mV ($\sim$25 pN) to 220 mV ($\sim$45 pN), but for three individual incorporation positions, the incorporation

rate either increased, decreased, or remained unchanged over this force range. Using this same dataset, we sought to determine the underlying rate constants that produced the diverse coupling between force and overall rate at these three positions.

Results: For every SPRNT read at the three individual incorporation sites (C1, U1 and U2 in Figure 66), we separated the position measurements into TEC states using a custom point-by-point level finding and alignment algorithm (see section methods). From this analysis, we calculated a mean dwell time ($\langle t \rangle$) for each of the TEC states (Pre, Post, and Hyper) and the branching probabilities out of Post (P_{Back}, $P_{Forward}$ and P_{Escape}) at each measured voltage (120, 150, 180 and 220 mV and) and at 10 μM [NTPs]. For position U1, we were unable to differentiate between Pre and Post at 120 mV, so measurements were limited to 150 mV and above. Figure 76 shows position *vs.* time plots for each position at the various applied voltages, where the different TEC states are colored based on the TEC state finding algorithm. For all three positions, we detected many visits to each of the TEC states, providing access to the full dwell time distributions for each state. The number of enzymes measured for each position and each voltage are tabulated in Figure 80.

We modeled the relationship between dwell time and force for each TEC state using the known thermodynamic relationship between rate and applied force. The rate is affected by the applied force so that ($k(F) = ke^{F\delta x/RT}$) where F is the applied force in the direction of RNAP motion, δx is the physical distance over which that force acts and RT is the temperature in units of energy. For the purposes of SPRNT, the force is directly proportional to the applied voltage, $F = qV/\delta x$ where q is an effective charge being moved through the electric field of magnitude $V/\delta x$. We can then write:

$k(V) = k(V = 0)e^{qV/RT} \equiv k(V = 0)e^{\alpha V}$ where α represents the coupling between the kinetics and voltage. Assuming that $q = 1e^{-}$ and using RT = 25 meV, we can calculate a physical upper-bound of $\alpha = 0.04mV^{-1}$

For any TEC state, the dwell time was equal to the inverse of the sum of the rate constants

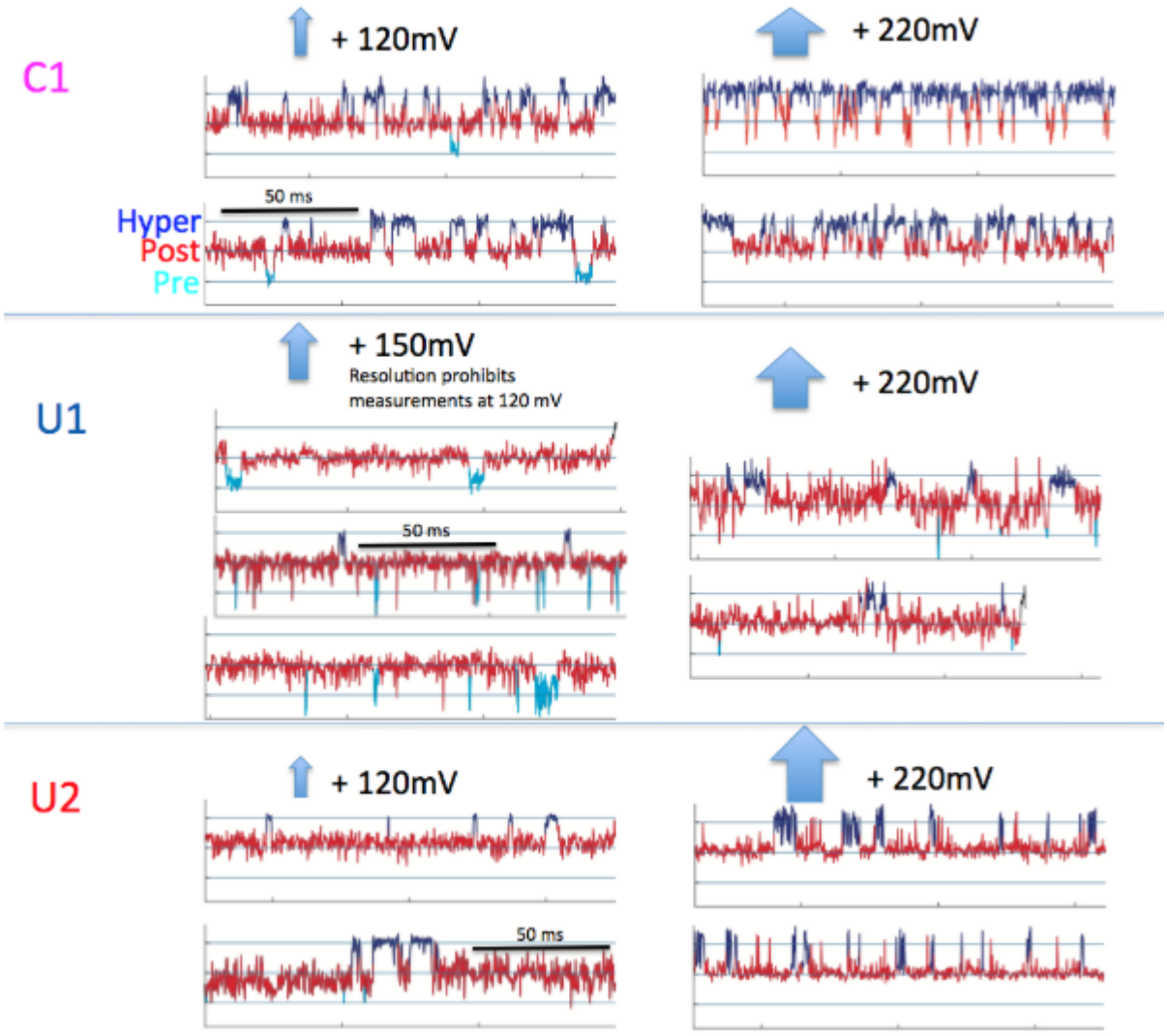


Figure 73: 150 ms of position *vs.* time data for each of the three incorporation positions (C1, U1 and U2) at low and high voltage. Each trace was acquired with a different individual RNAP enzyme. The position data was colored based on the assigned TEC states (cyan=pre, red=post, blue=hyper).

out of the state. For both Pre and Hyper only one kinetic path exits these states, meaning one rate constant controls the rate out of these state (k_1 for Pre and k_{-2} for Hyper), and the rate:voltage equations for these states are single exponential functions ($k_{pre}(F) = k_1(F)$ and $k_{pre}(F) = k_1 e^{\alpha V}$). As expected, for all three positions, the rate out of Pre increased with applied voltage as the rate constant out of Pre was assisted by the applied force. In addition, the rate out of Hyper decreased with applied voltage for all three positions, as the rate constant out of Hyper was opposed by the voltage. We fit these rate *vs.* voltage curves to the derived equations in Figure 75 and determined the rate constants at zero force as well as α (see Figure 77). The results of these fits yielded similar values for the k_{pre} for all three positions (Figure 77 column 1) as well as α (table 77 column 2). For all three positions, the zero-force rate constants out of Pre (k_1) were smaller than the rate constants out of Hyper (k_{-2}), which exhibited a wider range of values. In addition, the α values for the Hyper state varied significantly

For the Post state, the relationship between applied voltage and dwell time was more complex. For two of the three positions (C1 and U2), the rate out of Post (k_{Post}) increased with applied voltage, while k_{Post} decreased with applied voltage at the third position (U1) (Column 2 in Figure 76 A,B,C). For all three positions, the probability of a backwards transition from Post to Pre (P_{Back}) decreased with applied voltage, but the range of P_{Back} was much lower for C1 and U2 ($\sim$5% to $\sim$15%) compared to U1 ($\sim$45% to $\sim$65%)

There were three rate constants exiting Post ($k_{-1}(F)$, $k_2(F)$ and $k_{on}[NTP]$), so the overall Post rate was governed by the sum of these three rate constants. The rate forward into Hyper ($k_2(V) = k_2 e^{\alpha V}$) and Backward into Pre ($k_{-1}(V) = k_{-1} e^{-\alpha V}$) were both force-dependent, while the rate of NTP binding ($k_{on}[NTP]$) should be independent. We define the the relationship between k_{Post} and voltage:

$$k_{Post}(V) = k_2 e^{\alpha V} + k_{-1} e^{-\alpha V} + k_{on}[NTP].$$

The probability of a backwards transition out of Post was governed by the relative values

of the rate constants out of Post:

$$P_{Back}(V) = k_{-1}e^{-\alpha V}/(k_2 e^{\alpha V} + k_{-1}e^{-\alpha V} + k_{on}[NTP]).$$

In this model, we assumed that the rate of NTP catalysis (k_{cat}) was much larger than the rate of NTP unbinding (k_{off}). Under this assumption, every time an NTP binds in Post, the NTP is incorporated and the RNAP escapes to the next NTP incorporation cycle. Therefore, by measuring the probability of escape from Post (P_{escape}), we can estimate the NTP binding rate (k_{on}) using P_{escape} and k_{Post} ($P_{escape} = 1/num_{Post}$). We calculated the k_{on} at each voltage at each position and took the average of these measurements (Last column of the Table in Figure 77).

We performed simultaneous fits using the equations for P_{Back} and k_{Post} across applied voltages to determine the other rate constants out of Post (k_{-1}, k_2) and the force:voltage coupling factors (α) for each Position. These values are tabulated in Figure 77. Both the absolute and relative values of k_{-1} and k_2 varied significantly between the three positions (Figure 77) In addition, even though both U1 and U2 positions corresponded to UTP incorporation, we measured very different values for k_{on} at these two positions.

We used the derived rate constants and α values to estimate the average incorporation rate at each position. We calculated the expected number of visits to each state based on the rate constants (Figure 79). The incorporation time was estimated based on the number of visits to each state and the dwell time of each state. We plotted the expected incorporation rate for each position across applied voltages.

Discussion: These results are the first measurements of the rate constants during transcription elongation at individual incorporation positions. The derived rate constants help describe the diverse set of kinetic scenarios that led to the different couplings between incorporation rate and applied force across the three transcription positions. For both C1 and U2, the rate constants for transitions into Hyper at F=0 ($\sim$61 and $\sim$3.2 s^{-1}) were larger than the rate constant for transitions into Pre at F=0 ($\sim$36 and $\sim$4.1 s^{-1}) For both of these positions,

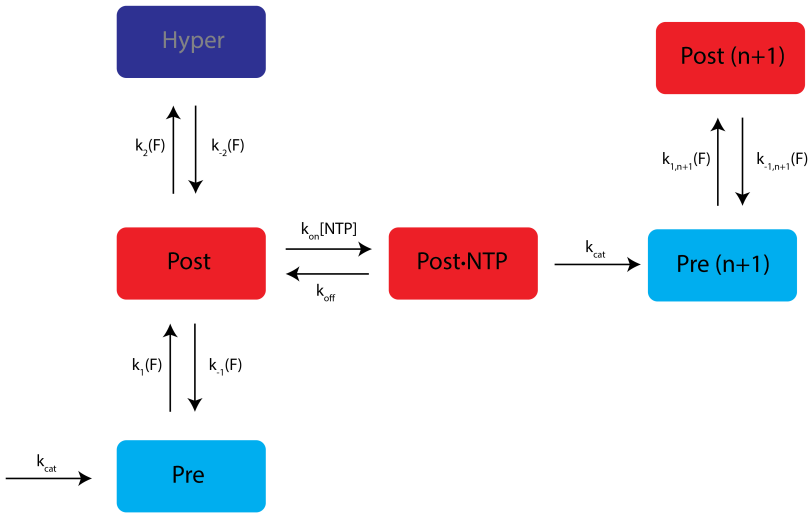


Figure 74: Schematic of 1.5 cycles of transcription elongation, where the TEC can oscillate between Pre, Post, and Hyper prior to NTP binding in Post. Vertical transitions represent steps in the reaction that coincide with physical movement of the RNAP along the DNA and are dependent on force. Horizontal transitions are not dependent on force. This model is simplified, and other arrows could potentially be added to this kinetic schematic to describe the data.

increasing the applied voltage increased the rate out of Post, resulting in a shorter dwell time in Post. As a result, the probability of NTP binding during any Post visit decreased and overall incorporation rate decreased. This incorporation rate decrease was more significant for position C1, as the rate out of Hyper was smaller compared to position U2, resulting in an increase in time spent in Hyper at higher applied voltages and decreasing the relative time spent in Post. In contrast, for position U1, the rate constant for transition into Pre was much larger than the rate constant into Hyper. For this position, increasing the applied voltage caused a decrease in the rate out of Post, increasing the probability of NTP binding and speeding up the incorporation rate.

For all three of the positions, based on the predicted incorporation rates, there were regimes where increasing the applied force resulted in an increase in incorporation rate while other regimes resulted in a decrease, with a transition between the two regimes where the transcription rate does not change significantly. Interestingly, although the overall kinetics

$$k_{Post}(F) = k_{-1}(F) + k_2(F) + k_{on}[NTP]$$

$$k_{Post} = k_{-1}e^{-\alpha V} + k_2 e^{\alpha V} + k_{on}[NTP]$$

$$k_{Hyper}(F) = k_{-2}(F)$$

$$k_{Hyper} = k_{-2}e^{-\alpha V}$$

$$k_{Pre}(F) = k_1(F)$$

$$k_{Pre} = k_1 e^{\alpha V}$$

$$P_{back,\,Post}(F) = \frac{k_{-1}(F)}{k_{-1}(F) + k_2(F) + k_{on}[NTP]}$$

$$P_{back,\,Post} = \frac{k_{-1}e^{-\alpha V}}{k_{-1}e^{-\alpha V} + k_2 e^{\alpha V} + k_{on}[NTP]}$$

$$P_{escape}(F) = \frac{k_{on}[NTP]}{k_{-1}(F) + k_2(F) + k_{on}[NTP]}$$

$$P_{escape}(F) = \frac{k_{on}[NTP]}{k_{Post}(F)}$$

$$k_{on}[NTP] = P_{escape}(F) \bullet k_{Post}(F)$$

Figure 75: Equations governing the rate in each state and transition probabilities. For each state, the rate for each state (k_{Pre},k_{Post},k_{Hyper}, where k=1/t) is equal to the sum of the rates out of each state and can be written in terms of applied voltage, where α is a coupling constant between force and voltage. The probability that the enzyme went back from Post into Pre ($P_{Back,Post}$) is governed by the relative values of the three rate constants out of Post. In our derivations, we assumed that the rate of NTP catalysis (k_{cat}) was much larger than the rate of NTP unbinding (k_{off}). For this reason, we assume that every time an NTP binds, that NTP is incorporated and the TEC enters the next Pre (n+1) state. Thus, the probability of escaping the individual reaction cycle (P_{escape}) is equal to the relative value of the rate of NTP binding to the other rate constants out of Post.

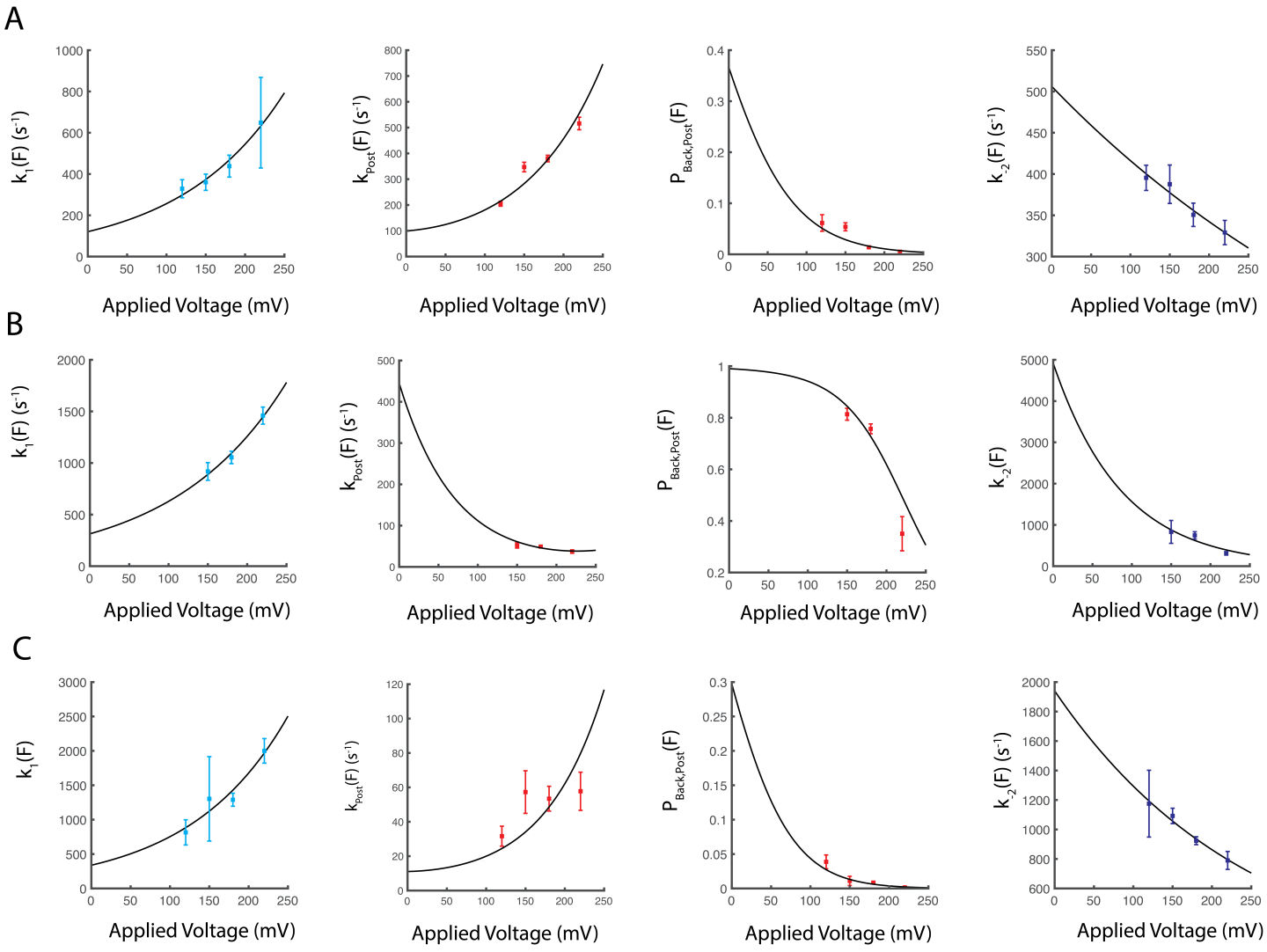


Figure 76: (A) Fits to the rate equations for Position C1. The rate out of Pre was measured at each voltage and fit to the equation for Pre in Figure 75 (column 1). The rate out of post (k_{Post}, column 2)) and the probability of a backwards transition from Post ($P_{Back,Post}$, column 3) were measured at each voltage and simultaneously fit to the equations for k_{Post} and $P_{Back,Post}$. The rate out of Hyper was measured at each voltage and fit to the equation for Hyper in Figure 75 (column 4). (B) Same as (A) except for Position U1 (C) Same as (A) except for Position U2

Position	k_1 (s^{-1})	α_{Pre}	k^{-1} (s^{-1})	k_2 (s^{-1})	α_{Post}	k_{-2} (s^{-1})	α_{hyper}	k_{on} (μM^{-1})
C1	120	.008	36	61	.010	506	0.002	0.2 ± 0.06
U1	314	.007	441	0.7	0.014	4918	0.012	0.4 ± 0.2
U2	338	.008	3.2	4.1	0.013	1942	0.004	0.07 ± 0 .01

Figure 77: Table of rate constants and α for each state calculated from the fits in Figure 76. The k_{on} for each site was calculated from the equation in Figure 75 at each voltage and an average value was calculated across voltages.

$$num_{post} = \frac{k_{-1} + k_2}{k_{on}[NTP]}$$

$$num_{pre} = \frac{k_{-1}}{k_{-1} + k_2} * num_{post}$$

$$num_{hyper} = \frac{k_2}{k_{-1} + k_2} * num_{post}$$

$$t_{inc} = (num_{pre} * t_{pre}) + (num_{post} * t_{post}) + (num_{hyper} * t_{hyper}) + t_{post}$$

Figure 78: The predicted average incorporation rate can be calculated from the rate constants in Figure 74. The number of visits to Post (num_{Post}) is governed by the probability of NTP binding in Post. The number of visits to Pre (num_{Pre}) and Hyper (num_{Hyper}) are governed by the relative value of the rate constants back and forward out of Post. The total average incorporation time (t_{inc}) can then be calculated from the average number of visits to each state and the dwell time in each state.

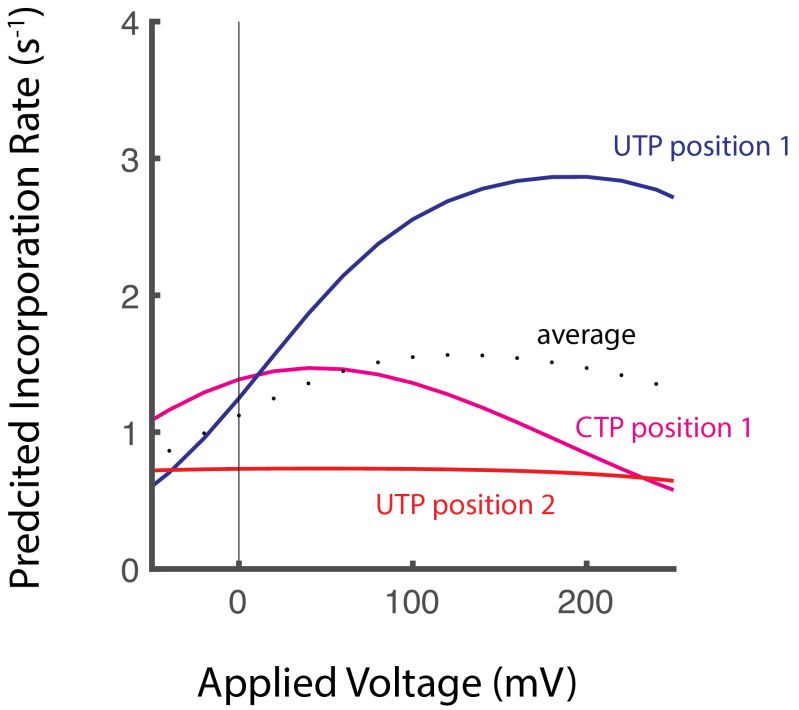


Figure 79: The predicted rate for the three incorporation positions across voltages was calculated using the incorporation time model derived from Fgiure 78 with the rate constants in Figure 77 at each position.

were very distinct for the three positions, the predicted incorporation rates at F=0 were similar (between ∼0.8 and 1.4 s^{-1})

These results also suggest that the simple model for transcription elongation shown in Figure 74 was insufficient to fully explain the data. For example, the calculated values for k_{on} were very different for the two U incorporation positions (∼0.4 $\mu M^{-1}s^{-1}$ for U1 and ∼0.07 $\mu M^{-1}s^{-1}$ for U2). We believe that the calculated value for k_{on} was actually a combination of multiple rate constants, some of which were not included in the model. First, the assumption that the NTP off rate (k_{off}) was small compared to k_{cat} may not be valid at certain transcription positions. Under these conditions, the equation for the probability of escape (P_{escape}) becomes much more complex. In addition, the NTP may unbind when the RNAP oscillates to Pre or Hyper after binding but before incorporation, and these pathways may have different rate constants than the transitions without the NTP present. Lastly, the NTP binding and unbinding pathway for RNAP in Post has previously been described with multiple steps, including possible entrance of the incorrect NTP type, as well as repositioning of an active site domain (the trigger loop) that stabilizes the NTP [54]. These steps cannot be detected directly with SPRNT and were hidden within the measurements of the Post dwell time. All of these different pathways combined to affect the measured k_{on} and the other rates out of Post. Investigation into the kinetics of the pathway at other [NTPs] and at different voltages will be required to tease out the details of this pathway, as well as the other unmeasured constants (k_{off}, k_{cat}).

While these measurements provide detailed insight into transcription elongation, they also highlight some of the limitations of this technique. Because the experiments were limited to voltages above 120 mV (∼25 pN), there were no data points close to zero applied voltage, and thus all of the exponential fits used to determine the rate constants were inherently error prone. In addition, we measured inconsistent α values for the different states and positions, particularly for the Hyper state, suggesting that the error in our measurements was large. It could also be the case that there were hidden sub states within the individual TEC states that were not force dependent (like structural rearrangement of other protein domains), but

Voltage (mV)	C1: n (fits)	U1: n (fits)	U2: n (fits)
120	87	0	19
150	79	49	57
180	152	125	79
220	40	40	23

Figure 80: Experimental counts for the data used to calculate the rate constants for each state at the individual incorporation positions (C1,U1 and U2). Each count represents an individual enzyme. The experimental counts were lower for these measurements compared to the measurements of average incorporation rate (Figure 71) as some of the traces were difficult to dissect for transitions between TEC state.

changed based on state or sequence, leading to inconsistent measurements for α. For this reason, while the calculated α values can be useful in determining rate constants, it was difficult to put much analysis into the absolute value of these constants for understanding the relationship between force and voltage. /par

Lastly, due to processivity concerns and the differences in resolution between positions along the scaffold, we only measured the rate constants for three incorporation positions. Although the average between these positions was a decent representation of the general behavior observed more broadly (Figure 70), more positions will need to be measured to understand the full range of relative rate constants. In addition, many more positions with varied DNA sequences will need to be measured to understand how particular sequence positions within the TEC modulate transcription behavior. A more systematic study will be required, perhaps where each nucleotide is changed independently within the TEC, will be needed to fully understand this complex system.

Methods:

Finding TEC state transitions:

To assign individual points in the ion current time series to different translocation po-

sitions, we used a hidden Markov model (HMM), whose true state transition probabilities were determined by the enzyme kinetics and the observed state probabilities were empirical distributions computed using a kernel density estimation method.

In brief, an HMM is a model describing a sequence of non-independent measurements, in which the state of the system changes in discrete steps, and at each step takes on a discrete "true state $\{\mathcal{S}_1, \mathcal{S}_2, ...\}$; $\mathcal{S}_n \in \mathbb{N}$. Each hidden state $\mathcal{S}_n$ is equipped with (1) an "emission" probability distribution $P_n(X); X \in \mathcal{O}$ over an observation space $\mathcal{O}$, which may be either continuous probability density function or a discrete probability mass function; and (2) a discrete set of transition probabilities $T_{ij} = P(\mathcal{S}_{n+1} = j|\mathcal{S}_n = i)$ describing the likelihood of transitioning from that state back to itself and to each other possible state in the next step of the model. To be classified as a Markov model, it is required that these transition probabilities depend exclusively on the present state $\mathcal{S}_n$, and conditioned upon the present state that they depend not at all on the history of states $\{\mathcal{S}_{n-1}, \mathcal{S}_{n-2}, ...\}$ that have been visited.

The model is used to interpret a Markov process, consisting of a sequence of measurements $\{X_1, X_2, ...\}$; $X_n \in \mathcal{O}$ on the observation space. The goal of the analysis is to assign each observation to the most likely true state that the system was in when the observation was made. This is done using either a *maximum a posteriori* (MAP) algorithm [94], which determines the most likely state for each individual observation, or a Viterbi algorithm [95], which finds the globally optimal simultaneous set of state assignments for all observations. In this work we opt for the latter, as in enzymology the entire time series of enzyme position measurements should be analyzed as a single experiment. This contrasts with, for example, the nanopore DNA sequencing problem, where we seek to minimize the per-base error rate, and a MAP algorithm is more appropriate [93].

A key part of any empirically informed hidden Markov model is the determination of the emission probabilities $P_n(X)$ and the transition probabilities T_{ij}. This is done through an expectation maximization (EM) algorithm, in which an initial guess for these parameters is

chosen, the observations are assigned states using the Viterbi or MAP algorithm, and the emission distributions and transition probabilities for each state are recalculated based on the data from the observations assigned to that state; this is then repeated until convergence.

For nanopore enzymology data, the observations $\{X\}$ are individual ion current points, the true states $\{\mathcal{S}\}$ are different DNA positions visited as a result of the activity of the anchoring enzyme, the emission probabilities P_n are the continuous distributions of ion currents expected from each state, and the transition probabilities T_{ij} are based on the kinetics of the enzyme.

The identities of the true states $\{\mathcal{S}\}$ are determined through consensus generation by eye, assisted with analysis software to highlight state transitions. Many reads are cross-compared to determine which distinct ion current states occur and in which order.

A representative sample of observations (ion current time series points) $\{X_1, X_2, ..., X_N\}$ is chosen for each true state, and is used to compute an empirical distribution estimating the emission probability distribution. We use an empirical distribution method because the distribution of ion current points in any individual state does not necessarily fit any one consistent functional form, and is frequently skewed or multimodal. The empirical distribution is found with a kernel density estimation method, where it is computed from a set of observations as

$$P(x|\{X_1, X_2, ..., X_N\}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-X_i)^2/2\sigma^2}. \quad (4)$$

A Gaussian distribution is used as the kernel to reflect the normally distributed error on individual ion current measurements. The standard deviation of the Gaussian kernel σ is a free parameter, which we set to 0.25 pA. This is large enough to ensure the smoothness and support of the estimated distribution given the number of samples, while it is sufficiently smaller than the typical width of the empirical distribution to resolve detail and prevent the kernel from washing out the larger-scale structure in the true distribution. Results are robust to choices of σ within a reasonable range satisfying these criteria.

In order to avoid rounding error, simplify computation, and to ensure support over the full probability space, we discretize the observation space in 150 equally spaced points x_i over the interval $x \in [\min_i X_i - 5\sigma,\ \max_i X_i + 5\sigma]$, and store the log probability at each point, $\log P(x_i)$.

During the execution of the Viterbi algorithm, when finding the log likelihood of a point in the interval $[\min_i X_i - 5\sigma,\ \max_i X_i + 5\sigma]$, we use a first order linear interpolation of the discretized log probability distribution described above. When finding the likelihood for a point outside that interval, the likelihood is dominated by the kernel contribution from the nearest point, and we approximate it simply as that probability, i.e. for a point $x < \min_i X_i - 5\sigma$, we use $\log P(x) = -\log \sqrt{2\pi} N\sigma - (x - \min_i X_i)^2/2\sigma^2$, and for a point $x > \max_i X_i + 5\sigma$, we use the same distribution but replacing $\min_i X_i$ with $\max_i X_i$.

Initial estimates for the transition probabilities T_{ij} are obtained simply by counting the number of transitions from each state to each other state in a representative set of by-eye classified observations.

With these inital guesses, we carry out the EM step, in which we solve the HMM with the Viterbi algorithm and re-calculate T_{ij} and P_n, now using the algorithm-assigned state identities in place of those selected by eye, repeating until convergence. The trained HMM can then be used to analyze new time series and identify the points in those time series where the enzyme steps and moves the DNA.

$\langle \mathbf{t} \rangle$: For every visit to each TEC state across all reads, we calculated the dwell time and took the mean value of all dwell times for each position ($\langle \mathrm{t}_{Pre} \rangle$, $\langle \mathrm{t}_{Post} \rangle$, $\langle \mathrm{t}_{Hyper} \rangle$). In order to properly estimate the error, we calculated the average dwell time for each read, and took the standard error in the mean (s.d.m.) of these measurments $= \sigma/\sqrt{n}$ The rate out of each state was equal to the inverse of the dwell time for that state: ($\mathrm{k}_{pre} = 1/\langle t_{Pre} \rangle$)

$\mathbf{P}_{Back}$: Across all reads, we measured the total number of transitions from Post into Pre (N_{Pre}) and Hyper (N_{Hyper}). $P_{Back} = N_{Pre}/(N_{Pre} + N_{Hyper} + N_{Reads})$ We calculated the error

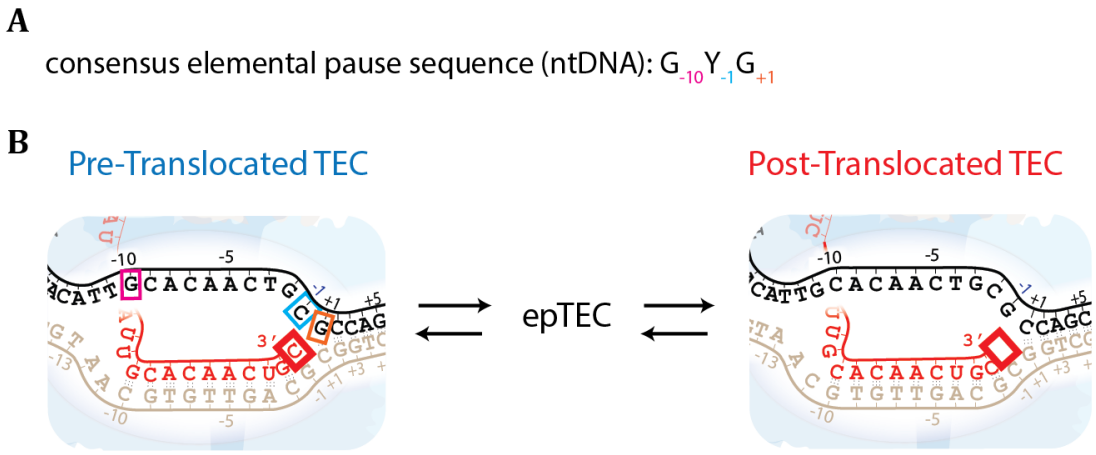


Figure 81: (A) Consensus pause sequence for *e. Coli* RNAP, where +1,-1, and +10 correspond to positions in the ntDNA in the Pre TEC (Y = C or T). According to previous studies, ∼16% of sequences with the consensus pause sequence cause pausing [39]. (B) Schematic of TEC at the yrbL pause sequence (a specific sequence with a 1:1 match of the G_{-10},Y_{-1},G_{+1}).

in P_{Back} using the binomial distribution:

$$\sigma_{P_{Back}} = \sqrt{P_{Pre} * (1 - P_{Pre})/(N_{Pre} + N_{Hyper} + N_{reads})}$$

$$P_{Forward} = N_{Hyper}/(N_{Pre} + N_{Hyper} + N_{Reads})$$
$$\sigma_{P_{Hyper}} = \sqrt{P_{Hyper} * (1 - P_{Hyper})/(N_{Pre} + N_{Hyper} + N_{reads})}$$

$$P_{Escape} = N_{reads}/(N_{Pre} + N_{Hyper} + N_{Reads})$$
$$\sigma_{P_{Escape}} = \sqrt{P_{Escape} * (1 - P_{Escape}/(N_{Pre} + N_{Hyper} + N_{reads})}$$

0.4.7 Elemental Pausing with SPRNT

Introduction: RNAP pausing, where the enzyme briefly stalls during transcription elongation, is an important part of the transcription pathway. We detailed the background of RNAP pausing in section 0.1.4.4. In short, many of the DNA sequences that cause pausing of *E. coli* RNAP have been determined and compared, leading to the consensus elemental pause sequence [39], [74]. The pause sequence corresponds to a G at -10, Y (C or T) at -1,

and G at +1 along the ntDNA strand (Figure 81). Pausing occurs at this sequence before G incorporation. While the kinetics of pausing have been investigated with other single-molecule techniques, direct detection of TEC pause states and measurements of dynamics during pausing are lacking.

Structural data of the *his* pause (a sequence containing the elemental pause sequence) suggests that a unique TEC state exists between Pre and Post at this sequence. During this state, termed the 'elemental pause state' (epTEC), the enzyme asymmetrically translocates with respect to the DNA and RNA (see detailed discussion in section 0.1.4.6). Although this state has been corroborated in multiple studies [79] [80], its precise role in pausing is not understood and definitive evidence outside of structural snapshots has not been obtained.

We monitored the translocation behavior of RNAP when encountering the *yrbL* pause sequence (another version of the elemental pause that has been investigated in other studies [39]. In the results already presented here, we noticed a significant spike in $\langle T_{total} \rangle$ at transcription positions that corresponded to the TEC encountering the yrbL sequence (section 0.3.3). Next, we sought to investigate yrbL pausing at high resolution and compare the dynamics of pausing to regular transcription elongation with SPRNT.

Results: We tracked enzyme position during RNAP pausing at the yrbl pause sequence (Figure 81). The yrbl sequence contains the conserved RNAP elemental pause sequence (G-10,Y-1,G+1, where Y = C or T in the ntDNA). Prior to the incorporation of the +1 guanine, RNAP stalls and is hypothesized to enter an elemental paused TEC state (epTEC). At 1000 μM NTPs and 180 mV ($\sim$35 pN) (n= 112 RNAP enzymes), we observed oscillation between up to five distinct TEC states during pausing and an average total pause lifetime of 1.13 $\pm$ 0.18 s. Using the 17 nt distance correction, we identified the paused TEC states as Backtracked, Pre, Half (with an average position of 0.48 $\pm$ 0.08 nts between Pre and Post), Post, and Hyper. These measurements are the first observation of a state between Pre and Post (the epTEC) in single-molecule traces. We did observe Half states at other transcription positions outside the pause sequence, but both the brief duration of these states and the low

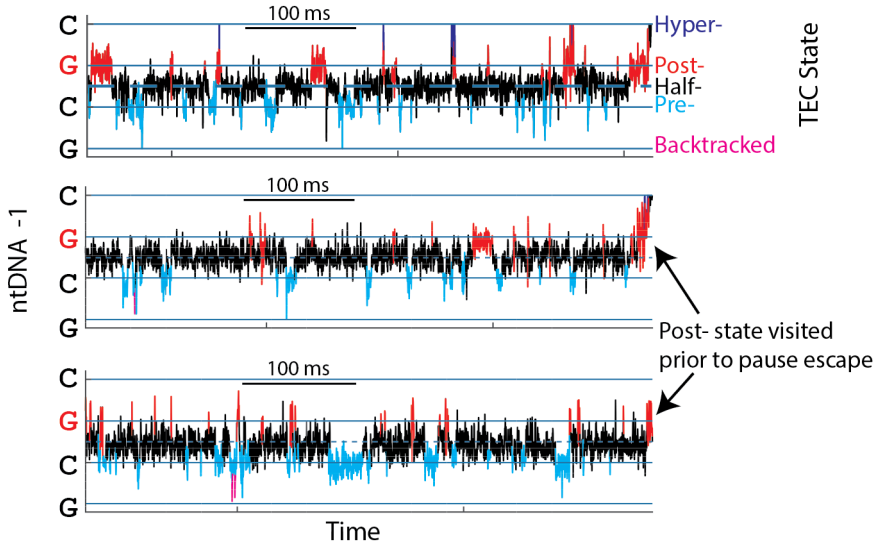


Figure 82: Three example traces of TEC state transitions during the last 250 ms of pausing at the yrbL sequence (180 mV, 1000 μM NTPs). Data plotted at 5 kHz, where the color of the positions measurements correspond to the different TEC states. Enzyme oscillates between up to five TEC states during pausing. Enzyme visited Post prior to pause escape.

probability of observation prohibited precise quantification of this behavior (Figure 63).

The Half state was the longest lived and most visited TEC state during pausing, and visits to Post were brief and rarely resulted in pause escape (1.7% $\pm$ 0.3%) (Figure 83). The average dwell time for the final visit to Post ($\langle$ $t_{Post,Final}$ $\rangle$) varied significantly from other visits to Post during pausing (Figure 84), whereas the final visit was indistinguishable from other visits for all other TEC states during pausing. In addition, the distribution of $\langle$ $t_{Post,Final}$ $\rangle$ was fit better by a multi-exponential function, whereas all other states were fit by a single-exponential, suggesting that multiple kinetic transitions occur during the final visit to Post (NTP binding, condensation (converting the TEC into the next Pre state (n+1)), and transition into the next Post state). At low [GTP] (10 μM), we measured an increased total pause lifetime (10.9 $\pm$ 2.1 s, n=47 enzymes), but the distribution of dwell times for each TEC state were indistinguishable compared to 1000 μM. (Figure 85). In fact, the only measurable change in kinetic parameters at low [GTP] was the probability of pause escape from Post (1.7% $\pm$ 0.3% vs 0.6% $\pm$ 0.1%). These measurements present a general rationale

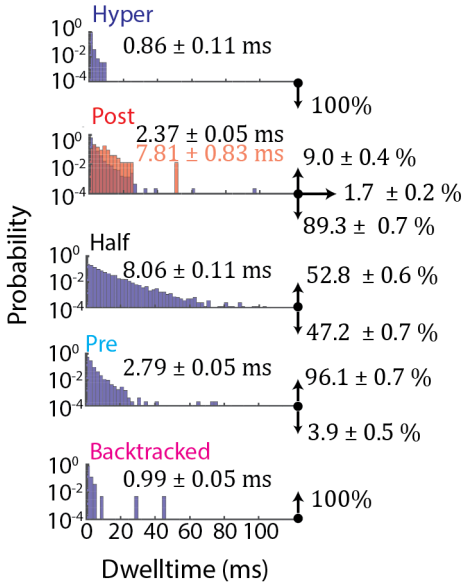


Figure 83: Dwell time histograms and branching ratios for each of the five TEC state during pausing (180 mV, 1000 μM). Data was combined for many enzymes (n=47). The dwell time distribution of only the last visit to Post prior to pause escape (orange panel 2) had a different shape distribution and longer dwell time than the distribution of all other visits (purple panel 2).

for long pauses at the yrbL sequence. The presence of a stable Half state (epTEC) blocks full translocation between Pre and Post, while the very brief duration of the Post state makes GTP binding and incorporation unlikely on any given visit, even at high [GTP].

In addition to the average behavior of RNAP at the yrbL sequence, we also observed and quantified significant static disorder during pausing (i.e. variation between the kinetics of individual RNAP core enzymes in a homogeneous population). For longer duration pauses, the Half state was visited many times prior to pause escape, allowing measurement of the Half state dwell time (t_{half}) distribution for individual enzymes (Figure 86)A). The distribution of t_{half} for individual enzymes varied significantly from the t_{Half} distribution for all enzymes combined. Individual enzyme distributions were each fit by a single-exponential function according to the rate constants dictating $\langle\ t_{Half}\ \rangle$ for that individual enzyme. For this

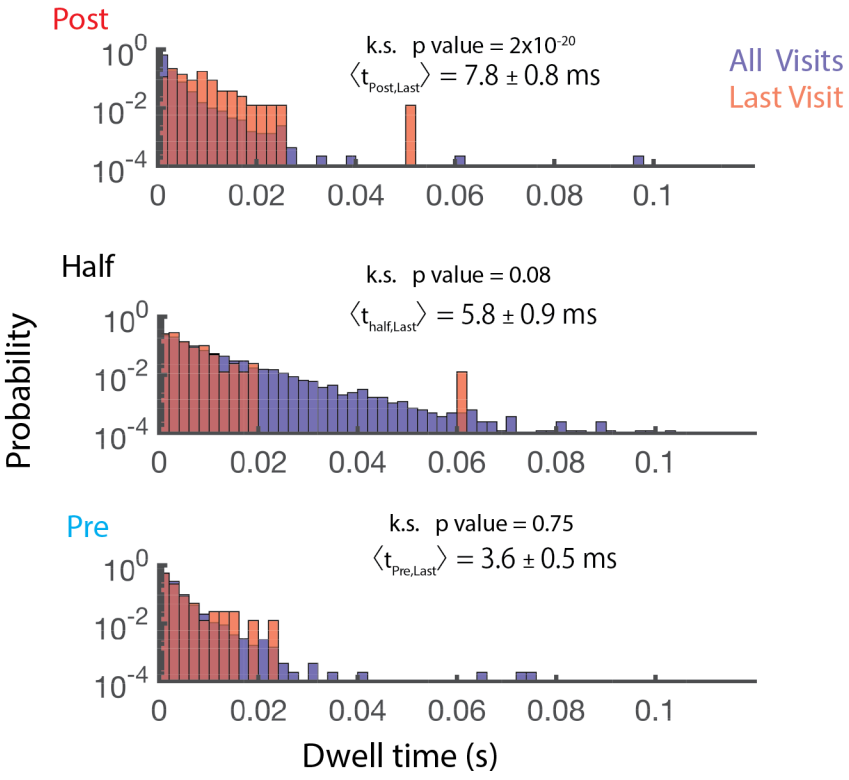


Figure 84: Dwell time histograms for the last visits (orange) and all visits (purple) for the three central TEC states during pausing (180 mV, 1000 μM). Data was combined for many enzymes (n=47). The k.s. values for each histogram evaluate the probability that the two datasets were acquired from the same underlying distribution.

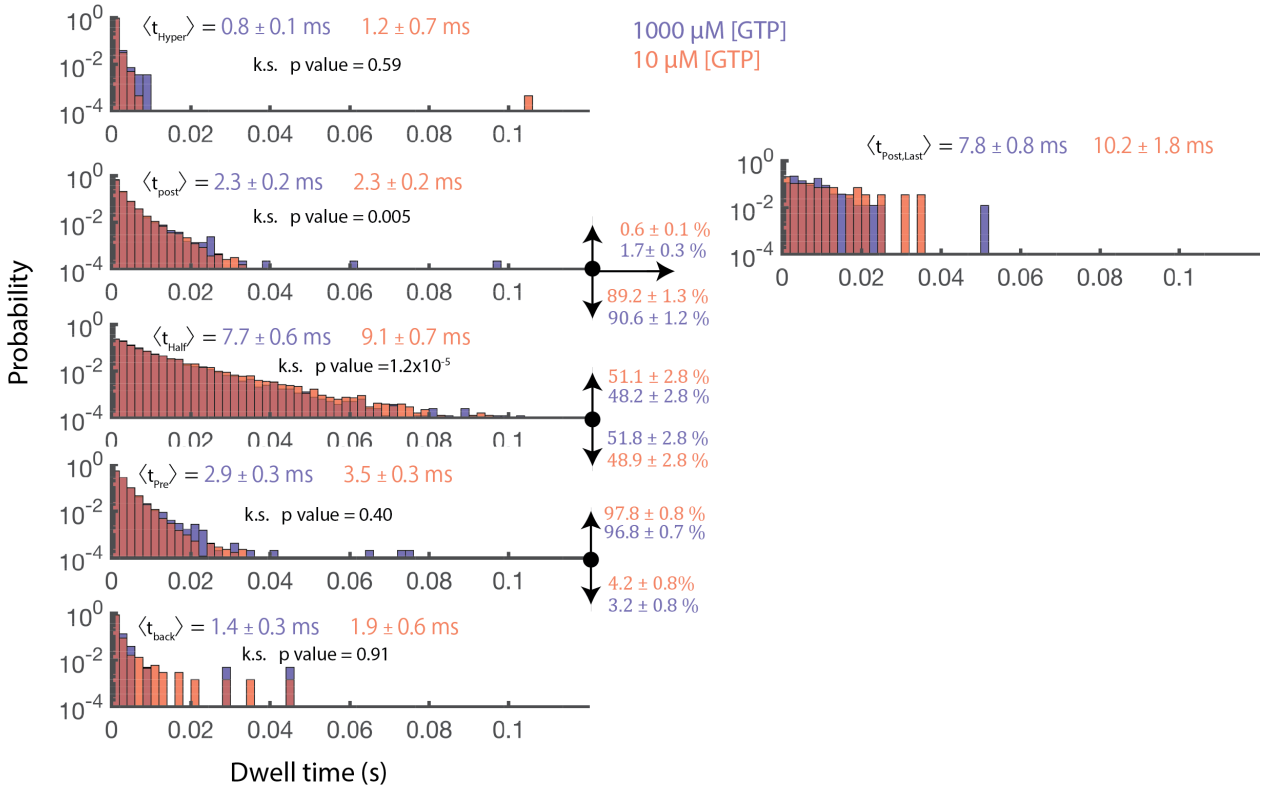


Figure 85: Dwell time histograms and branching ratios for each of the five TEC state during pausing, as well as the last visit to Post, at both 10 μM (orange) and 1000 μM (purple) [GTP] (180 mV). The k.s. values for each histogram evaluate the probability that the two datasets were acquired from the same underlying distribution.

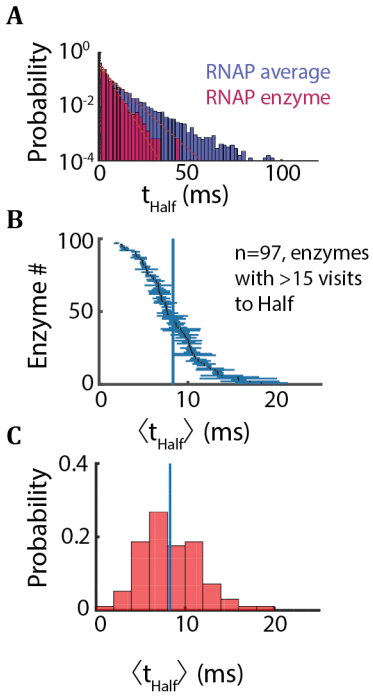


Figure 86: (A) t_{Half} histogram for an individual RNAP enzyme compared to combined histogram for all measured RNAP enzymes. Single-exponential fit to individual enzyme (red) and combined (cyan). (B) t_{Half} with s.d.m. for each RNAP enzyme with over 15 visits to Half (n = 97 enzymes). Average t_{Half} (blue line). (C) Histogram of t_{Half} for enzymes in (D) Average t_{Half} (blue line).

reason, the t_{Half} distribution for all enzymes combined is an overlay of many unique single-exponential distributions and appears stretched. Across 97 enzymes with more than 15 visits to the Half state, $\langle\, t_{Half} \,\rangle$ varied significantly (Figure 86)B) (Average $\langle\, t_{Half} \,\rangle$ was 8.3 ± 0.3 ms with a wide range from 2.0 ± 0.3 to 18.0 ± 3.0). The distribution of $\langle\, t_{Half} \,\rangle$ for this group of enzymes peaked around the average $\langle\, t_{Half} \,\rangle$ and did not contain identifiable subpeaks (Figure 86)C), suggesting that there is a continuum of rate constants for this population of RNAP enzymes rather than distinct subpopulations. The dwell time distributions for all other TEC states during pausing also exhibited significant static disorder.

In order to investigate the effect of applied force on pause dynamics, we varied the applied voltage between 120 mV and 220 mV ($\sim$+20 pN to $\sim$+42 pN) and tracked many RNAP enzymes during yrbL pausing. Clear identification of enzyme states during pausing was

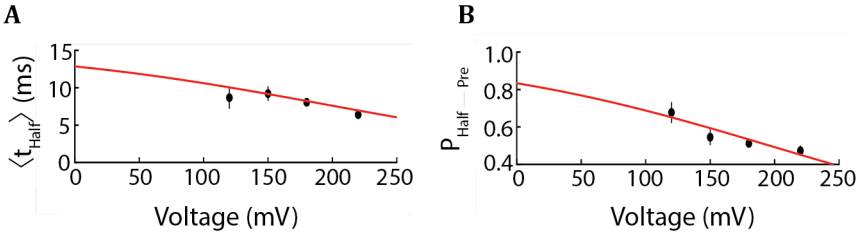


Figure 87: Average t_{Half} across a range of applied voltages (120 mV to 220 mV) (left panel). Average backward transition probability ($P_{HalftoPre}$) across a range applied voltages (120 mV to 220 mV). Simultaneous fits to t_{Half} and $P_{HalftoPre}$ vs applied voltage (red line).

prohibited below 120 mV. Both the dwell times and branching ratios for all TEC states during pausing varied significantly across the range of applied forces measured. For the epTEC specifically, both $\langle\ t_{Half}\ \rangle$ and the probability of returning to Pre from Half ($P_{Halftopre}$) decreased as the voltage increased (Figure 87). By simultaneously fitting the relationship between voltage and both $\langle\ t_{Half}\ \rangle$ and $P_{Halftopre}$, we calculated the underlying rate constants out of the half state to be k- $= 68 \pm 9\ s^{-1}$ and k+ $= 18 \pm 7\ s^{-1}$. Therefore, with no assisting force, we estimate that RNAP will spend significant time in the Half state during pausing at the yrbL elemental pause sequence ($\langle\ t_{Half}\ \rangle = 11.6 \pm 4.7$ ms , $P_{Halftopre} = 79 \pm 34$). The details of the modeling of RNAP pausing and implications of these results are discussed in further detail in section 0.4.7.1

Methods:

Finding TEC state transitions during pausing: We used the same state finding algorithm as in section 0.4.6.4 to differentiate between states during pausing. The beginning and end of a pause were recognized "by eye" and used to calculate the pause duration.

0.4.7.1 Modeling of RNAP Pausing

We measured 5 distinct observable states during pausing at the sequence-dependent pause site yrbL: Backtracked state, a Pre-translocated state, a Half-translocated state, a Post-translocated state and a Hyper translocated state. The dwell times and branching ratios be-

tween these states are each force-dependent, implying that transitions between these states are mechanical. The goal of this section is to develop a minimal kinetic model for translocation between these 5 observed states consistent with the data presented in the article.

Initial modeling: We observed that most transitions between observable states tend to occur sequentially between Backtracked, Pre, Half, Post and Hyper states, respectively. Combining this fact together with the knowledge that escape from the pause proceeds via the post-translocated state and that the chemistry of NTP incorporation occurs during the post-translocated state, the simplest possible kinetic model that we can develop is shown in Figure 88. In this model the RNAP toggles between the five observable states, rectified eventually by the binding and hydrolysis of the GTP molecule during the post-translocated state.

A comment on error analysis: In SPRNT, our observable parameters for a given enzyme state are the distribution of dwell-times, ρ, and the transition probability between observable states i and j (p_{itoj}), averaged over many RNAP molecules. The distribution for most of these states is a single-exponential distribution, so we can simply analyze the first moment $\langle t \rangle$. The errors on $\langle t \rangle$ and $p_{i \to j}$ are then given by the expressions:

$$\delta_{\langle t \rangle} = \langle t \rangle / \sqrt{N}$$
$$\rho_{p_{i \to j}} = \sqrt{p_{i \to j} * (1 - p_{i \to j})/N_i}$$

Where N is the number of measurements of a given state taken over all RNAP molecules. Because RNAP will ratchet between these states many times during a translocation event, N is large and therefore these errors become relatively small.

We found, however, that errors calculated using this method tended to be small compared to the variance between individual RNA polymerase traces (see Figures 86 and 91). Therefore, instead of calculating the error as above, we instead calculate dwell-times $\langle t \rangle_i$, and branching ratios $p_{i \to j}$ for each RNAP molecule that we measured and calculate the sam-

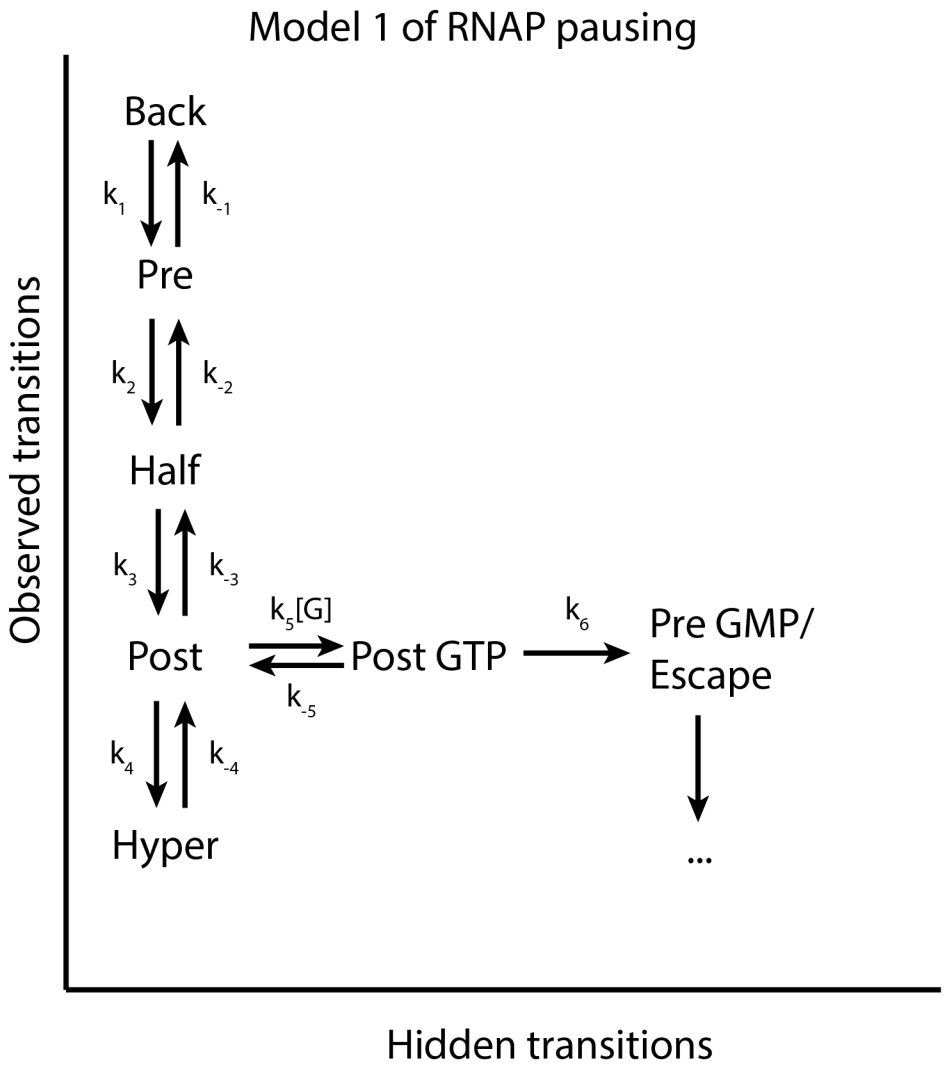


Figure 88: Model 1 of RNAP pausing at the yrbL pause sequence. The TEC transitions sequentially between enzyme states and GTP binding, unbinding and incroporation all occur in the Post state.

ple standard deviation in the mean of these measurements. This provides a more accurate representation of our uncertainty in key kinetic parameters. We discuss the limitations of this approach below.

Force-dependent rate constants: Any kinetics step that involves physical motion of the motor can be perturbed by the application of a mechanical force. In SPRNT, this force is applied by the voltage which establishes an electric field that pulls on the charged NAs. We assume force-dependent rate constants take the following form:

$$k(F) = k(F = 0)e^{F\delta x/RT}$$

where F is the applied force in the direction of RNAP motion, δx is the physical distance over which that force acts and RT is the temperature in units of energy. For the purposes of SPRNT, the force is directly proportional to the applied voltage, $F = qV/\delta x$ where q is an effective charge being moved through the electric field of magnitude $V/\delta x$. We can then write:

$$k(V) = k(V = 0)e^{q/RTV} \equiv k(V = 0)e^{(\alpha V)}$$

where α represents the coupling between the kinetics and voltage. Assuming that $q = 1e^-$ and using RT = 25 meV, we can calculate a physical upper-bound of $\alpha = 0.04mV^{-1}$

Analysis of the Half state: We analyzed the force dependence of the Half state by assuming that the Half state moves only to Pre and Post states. Because these steps involve physical motion of the RNAP, we assume that both k_{-2} and k_3 are force-dependent. We can then write the average dwell-time and branching ratio as:

$$\langle t \rangle_{Half} = 1/(k_{-2}(V = 0)e^{-\beta V} + k_3(V = 0)e^{-\gamma V})$$

$$p_{Half \to Pre} = k_{-2}(V = 0)e^{-\beta V}/(k_{-2}(V = 0)e^{-\beta V} + k_3(V = 0)e^{-\gamma V})$$

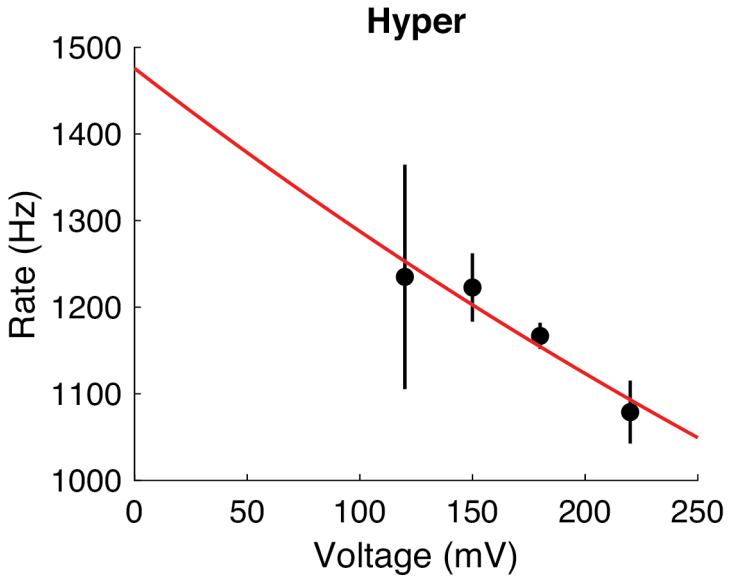


Figure 89: The rate out of Hyper decreased with applied voltage. A single-exponential fit (red line) was used to determine the underlying rate constant and voltage kinetic coupling factor.

These equations contain 4 free parameters $(k_{-2}(V=0), k_3(V=0), \beta, \gamma)$ which can be determined through fitting of the experimental data to these expressions. We measured branching ratios and dwell-times at four voltages (120 mV, 150 mV, 180 mV, 220 mV), giving us 8 measurements of dwell-time/branching ratio. Fits were performed by minimizing the global χ^2_v with parameters constrained to be larger than 0. From fitting to these expressions we find that $k_3(V=0) = 18 \pm 7 s^{-1}, \gamma = 0.007 \pm 0.002 mV^{-1}, \beta = 0.0002 \pm 0.0006 mV^{-1}$ (Figure 87)

Hyper-translocated: Because the hyper-translocated state almost always transitions to the post translocated state, we have:

$$\langle t \rangle_{hyper} = 1/k_{-4}(V=0)e^{-\delta V}$$

Fitting to the data we find that $\delta = 0.0015 \pm 0.0005 mV^{-1}$ and $k_{-4}(V=0) = 1476 \pm 144 s^{-1}$

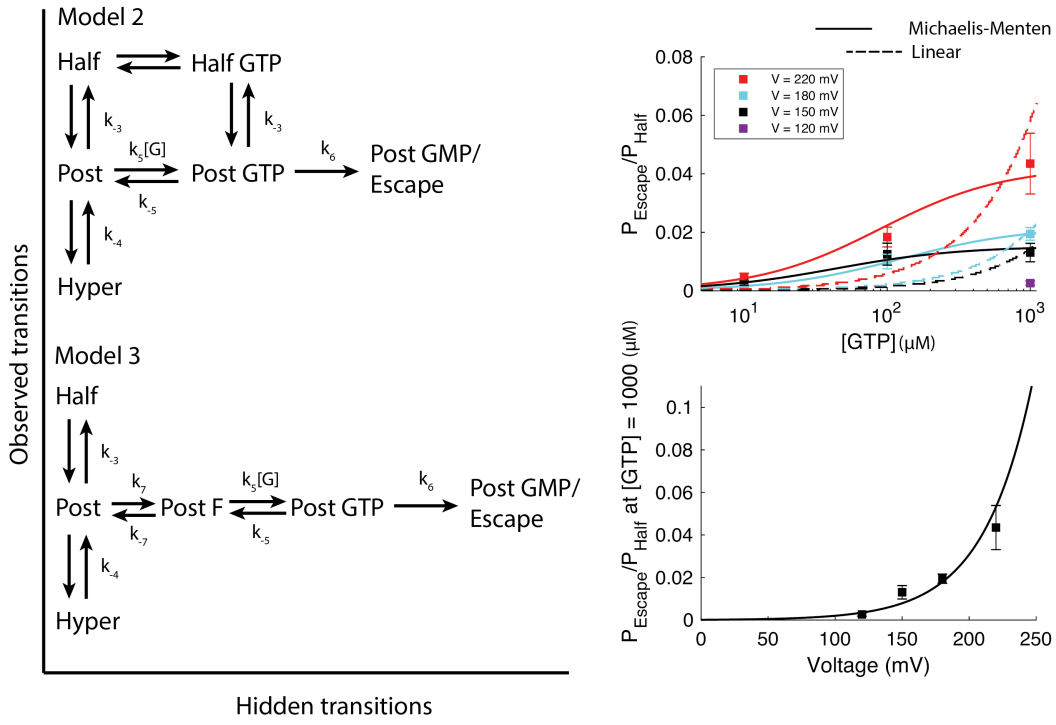


Figure 90: The RNAP pausing model can be expanded to include other kinetic transitions. In model 2, GTP can bind in either Half or Post. In Model 3, an extra step in Post occurs prior to GTP binding. We calculated p_{escape}/p_{Half} at each voltage and each [GTP] (10, 100 and 1000 $/mu$M) The ratio of p_{escape}/p_{Half} across different [GTP] was fit better by a michaelis-menten model than a linear model at every voltage, suggesting that either model 2 or model 3 describes the data better than model 1. For a single [GTP], p_{escape}/p_{Half} increased exponentially with voltage.

Escape from the Post-translocated state: To analyze escape from the translocation we start by calculating the probability that the enzyme escapes the pause as in our intiial model of pausing (Figure 88. Our initial model of RNAP pausing says that from the Post-translocated state, a GTP molecule binds to the RNAP in the Post-translocated state, from which GTP is hydrolyzed and incorporated into the growing RNA chain. Assuming this simplistic model, we can calculate the relative probability that the enzyme escapes the pause compared to moving backwards to the Half-translocated state [9]. For model 1 (Figure 88) we derive:

$$p_{escape}/p_{Half} = k_5[GTP](k_6/(k_{-3} + k_6 + k_5))$$

This expression is linear in the [GTP]. In Figure 90 we plot the resulting fit for our data at each of the three measured voltages, along with linear fits to the data. We find the reduced Chi-square (χ^2_v) to be 13, reflecting a poor fit. Indeed, the data appear to saturate at large [GTP], with the ratio p_{escape}/p_{Half} changing substantially between [GTP] = 10 μM and 100 μM, but very little between 100 μM and 1000 μM at each tested voltage. Motivated by these results, we proposed two slight variations on this model: one in which the RNAP can translocate between Half- and Post- states even with a bound GTP (model 2), and one in which some mechanism prevents GTP binding, before the GTP is finally bound and translocation proceeds. This state could be a folded state of the trigger loop, for example, but here the arrows are hypothesized based solely on the non-linear nature of these data to understand potential mechanisms for this saturating effect. In both of these models we calculate a Michaelis-Menten dependence of p_{escape}/p_{Half} on the [GTP]. For model 2, in which the GTP bound RNAP can still oscillate between Post and Half states, we calculate:

$$p_{escape}/p_{Half} = k_6k_5[GTP]/(k_{-3}k_5[GTP] + k_4k_{-3} + k_6k_{-3} + k_{-5}k_{-3} + k^2_{-3})$$

Fitting the data to these results we find $\chi^2_v \sim 0.7$, reflecting a much better fit to the data. In addition, at saturating [GTP] we find that the ratio p_{escape}/p_{Half} for this model to be exponentially dependent on the voltage

$(p_{escape}/p_{Half}) \sim (k_6/k_{-3}(F)) \propto e^{\gamma V}$

Similar expressions can be derived for model 3. We fit the data at [GTP] = 1000 μM and find the data are well described by a single-exponential, with $\gamma \sim 0.025 \pm 0.010 mV^{-1}$. A large assisting force tends to keep the enzyme in the Post-translocated more often, allowing for a higher escape probability.

Limitations of this analysis: There are several factors that limited the analysis performed here which can be improved upon in future experiments and data analysis

1) We mentioned previously that each individual RNAP molecule has its own set of kinetic rate constants. Here we attempted to take this into account by analyzing the distribution of rate constants and characterizing the spread of those data. This is not, strictly speaking, a correct approach. A proper treatment would be to include with each rate constant a distribution function that characterizes the distribution of k_i values for each enzyme, and to minimize a cost function over these distributions using Bayesian methods. Interestingly, the distribution of time constants $\tau_i = 1/k_i$, appears to nearly follow a gaussian distribution (Figure 91), which may simplify such a calculation. Such an approach could be performed globally on the data for all states using Hamiltonian Monte Carlo. This advanced statistical approach is beyond the scope of this text, and the fits performed above can be considered as rough estimates of the kinetic parameters. Importantly, we find here that the force-dependent coupling parameters are always less than 0.04 mV^{-1} which is a physical upper limit.

2) Below $\sim$ 150 mV transitions between the Pre and Back states became almost impossible to distinguish from one another, preventing fitting for those states being analyzed as we did for the Half- and Hyper- states. Because in experiments done with a $'$backwards MspA pore$'$ the RNAP sits so close to the $'$measure sequence$'$, it is difficult to change the measure sequence to optimize the SPRNT signal without disrupting the yrbl pause element. An equivalent set of experiments done on a $'$forwards pore$'$ would give more freedom to customize the DNA

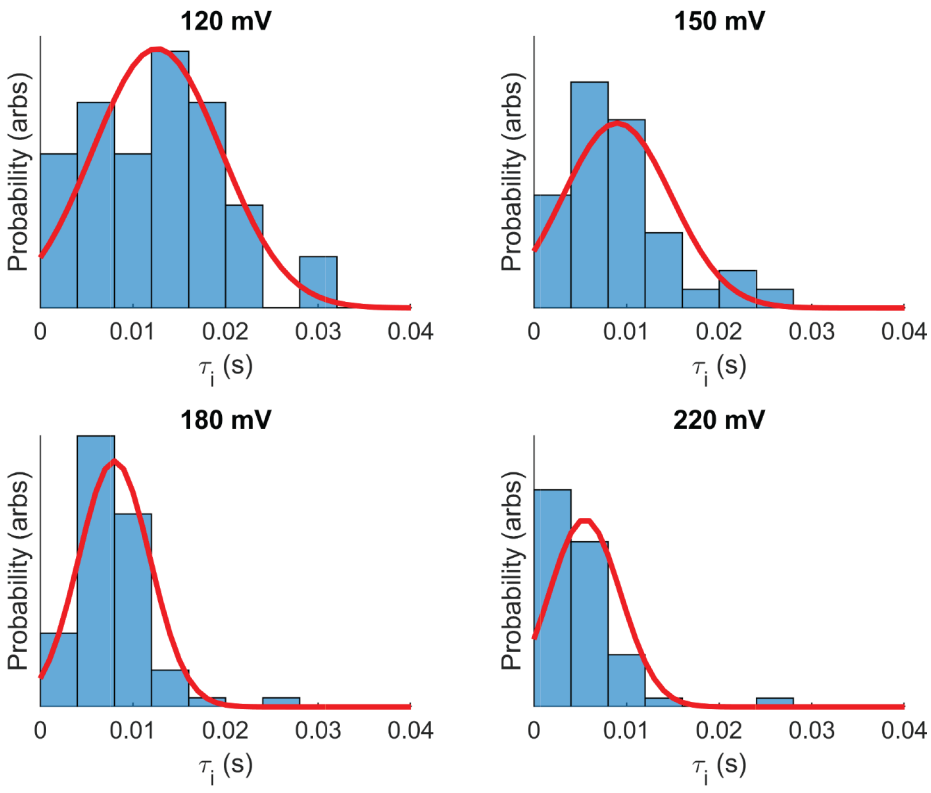


Figure 91: Histograms of Half state dwell times for individual enzymes at each voltage. While the average dwell time decreased with voltage, there was significant variation between enzymes at each voltage. Gaussian fits were performed (red line) to each of the dwell time distributions.

sequence, potentially even using abasics, to produce a large-contrast ion-current signal which would allow reduction of the voltage to lower values that we could reach in these experiments. The ability to resolve more of these states at low force will lead to more reliable modeling and extrapolations to the low-force regime.

0.4.8 Implications for detecting an intermediate state with SPRNT

In our SPRNT measurements of RNAP, we measure the position of the upstream tDNA outside of the enzyme. This measurement is simultaneously sensitive to 1) movement of the upstream tDNA relative to the edge of RNAP and 2) movement in domains of RNAP which form the RNAP:MspA interaction (e.g. what part of the enzyme sits on the pore). Movement of these domains would change the position of the tDNA in the pore and could not be distinguished from shifts of tDNA:RNAP contacts. Based on the structural data discussed in *Background*, measuring upstream tDNA position changes between the Pre and Post state are straightforward. The nucleic-acid:RNAP interactions are all shifted by 1 position, without conformational change of RNAP around the upstream region, so the tDNA in the measure sequence is shifted by 1 position [79] [80].

The measurement of the ePEC with SPRNT is less straightforward. Positioning changes of the DNA within the enzyme can only be detected if they affect the register of the tDNA at the RNAP: tDNA interactions at the upstream edge of the protein. In the proposed ePEC structures, there is a positional shift that propagates all the way to the upstream end of the tDNA /RNA hybrid (position -10 in the tDNA). The last direct tDNA:RNAP backbone interaction occurs between -10/-11 tDNA with arginine 259 (part of the B lid), so it could, in theory, be affected by this positional shift and change the positioning of the tDNA measured in the nanopore (the position of arginine 259 is not discussed in either of the cry-EM publications [79] [80]).

However, the structures predict small conformational changes in the regions affecting MspA:RNAP interaction as well. The upstream DNA exit through the protein occurs in between elements of the swivel module (specifically between the protrusion, clamp and SI2/flap

modules, [79]). In addition, the B lid (containing arginine 259) is associated with the opening of the RNA exit channel. Movement of any of these domains should affect the positioning of the tDNA in the nanopore. The exact nature of and magnitude of these conformational changes (or if these changes happen at all prior to his hairpin formation) seem to be under debate.

In summary, the measurement of a Half state with SPRNT could be due to an actual half position shift of the tDNA within the enzyme, or conformational shifts of the RNAP domains in contact with MspA, or most likely, some combination of the two. Directly comparing the different structures can be difficult as the expected distance change between Pre and Half is expected to be about 3 angstroms and the resolution of the cryo-EM ePEC from [80] is 5.5 angstroms.

0.4.9 Effect of Core-Recognition-Element on pausing with SPRNT

Introduction: During pausing, *E. coli* RNAP makes sequence-specific interactions with the downstream part of ntDNA ("core recognition element," CRE). Specifically, while in the Post state, the ntDNA guanine at the downstream edge of the transcription bubble (position -1 in the ntDNA) forms a hydrogen bond with RNAP β residue D446. This interaction plays a significant role in pause escape. Removal of the RNAP-G_{CRE} interaction using mutant RNAP D446A has previously been shown to enhance pausing at the yrbL sequence, presumably by destabilizing the Post state [39].

Structural data of the ePEC is inconclusive on the role of the RNAP-G_{CRE} in the Half state. There is disagreement between the positioning of the downstream ntDNA nucleotides between the crystal structure [73] and cryo-EM structures [79] [80] of the ePEC (see background). In the cryo-EM structures of the ePEC, the downstream ntDNA guanine was still hybridized, resembling the Pre state. This would mean that the RNAP-GCRE interaction is not made during the ePEC. However, in the earlier crystal structure, the ntDNA G was unpaired, even though the tDNA still fails to translocate, and therefore RNAP-GCRE could happen in Post.

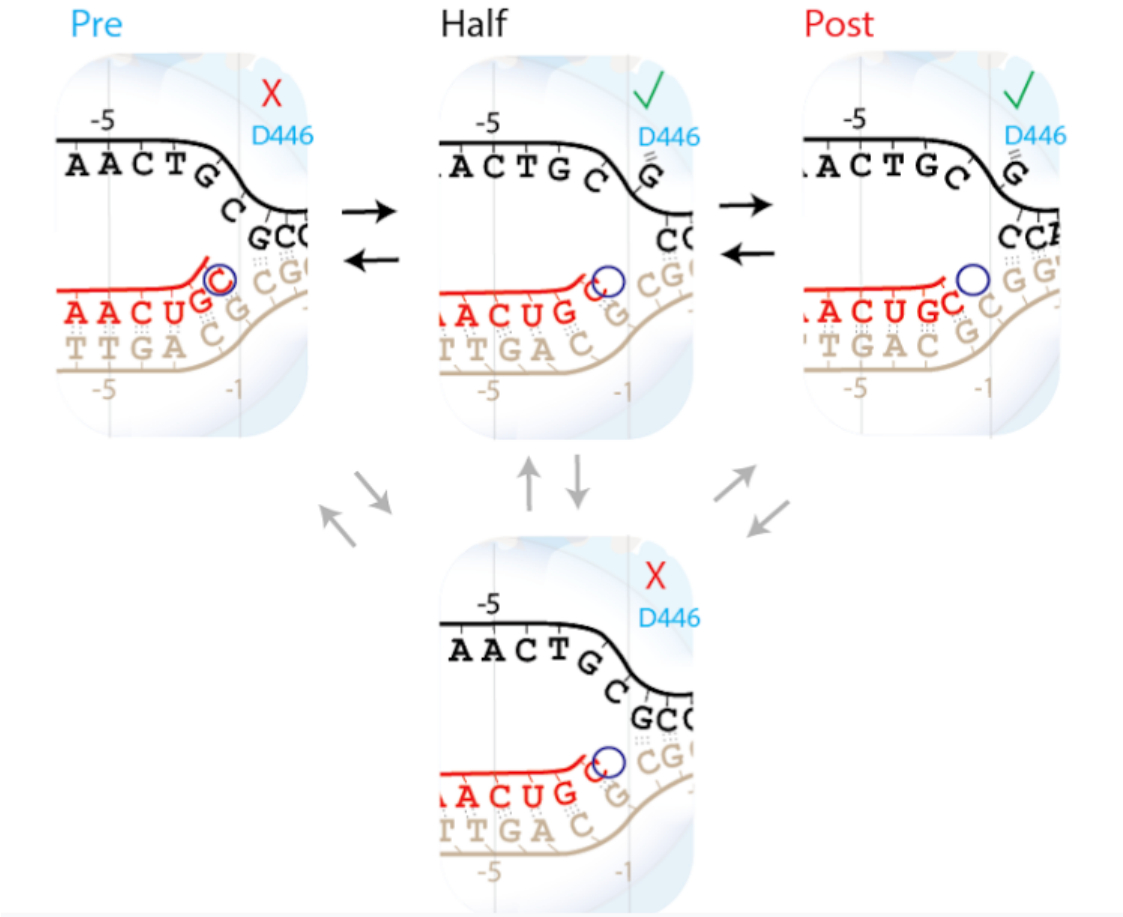


Figure 92: Structural data provides a model for pausing at the elemental pause sequence. Translocation of the RNA, but lagging of the tDNA, results in a tilted hybrid in the Half state. Substrate loading and GTP incorporation cannot occur until full translocation of the tDNA. Models differ on the location of the ntDNA during the half state. The RNAP-Gcre interaction is made in Pre and Post, but would only be made during Half if the ntDNA is translocated

Results: We monitored transcription and pausing at the yrbL pause sequence with mutant RNAP D446A enzyme using SPRNT. Average pause lifetime compared to wt increased significantly at both 1000 μM (3.51 $\pm$ 0.82 s at 180 mV, n=98) and 10 μM [GTP] (37.68 $\pm$ 8.21 s at 180 mV, n=40) (Figure 93). The mean values of pause lifetime were affected significantly by very long tails in the distribution of lifetimes, further demonstrating significant static disorder amongst enzyme populations (median pause lifetime for RNAP D446A was 0.8 $\pm$ 0.2 s at 1000 μM [GTP] and 14.8 $\pm$ 7.6 s at 10 μM [GTP] compared to 0.4 $\pm$ 0.1 s at 1000 μM and 5.1 $\pm$ 2.2 s at 10 μM for RNAP wt (errors were calculated by bootstrapping (see methods)). The shape of the distribution of lifetimes between D446A and wt vary significantly prior to the tail region, suggesting very different average pausing rate constants between the two enzymes.

Much of the observed pausing behavior was conserved between the wt and D446A. RNAP D446A visited the same five TEC states during pausing (Back, Pre, Half, Post and Hyper) (Figure 94). The last visit to the Post state prior to pause escape varied significantly from all other visits to Post for RNAP D446A, while the final visits to the other states were indistinguishable from all other visits, just as with the wt (Figure 95). The distribution of dwell times for the final post visit was better fit by a multi-exponential function. Lowering [GTP] also had a similar effect on pause kinetics for the mutant enzyme as the wt. Between [GTP] = 10 μM and [GTP] = 1000 μM, the distribution of dwell times for all of the TEC states were unchanged (Figure 96). Again, the only parameter that depended on [GTP] was the probability of escaping the pause from Post (0.6% at 1000 μM vs 0.08% at 10 μM).

Other parameters varied significantly between wt and D446A during pausing. The average dwell time in both Half and Post decreased significantly compared to wild type, while the dwell time in the Pre state was unchanged (Figure 99). RNAP D446A was also more likely to return to Pre from Half than wt. For these measurements, we combined the data from 10 μM and 1000 μM [GTP] in order to average over more enzymes. Most of the error is dominated by variation between enzymes in the same population. We, therefore, calculated the average by defining a dwell time for each enzyme with over 5 visits to each state (n=70

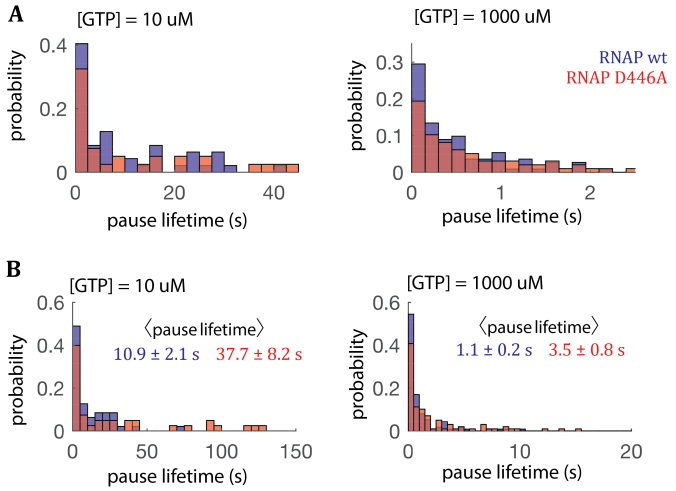


Figure 93: (A) distribution of pause lifetimes for 10 μM (left panel) and 1000 μM (right panel) GTP at 180 mV. Pause lifetime is increased in both [GTPs] with RNAP D446A (n=98 RNAP D446A enzymes at 1000 μM, n=40 RNAP D446A enzymes at 10 μM). (B) Same as (A) with extended x axis to show long tail of pause lifetime distributions. The average pause lifetime was heavily influenced by very long pauses. Median pause lifetime for RNAP D446A was 0.79 0.22 s at 1000 M [GTP] and 14.83 7.58 s at 10 M [GTP] compared to 0.39 0.08 s at 1000 M and 5.07 2.21 s at 10 M for RNAP wt (errors were calculated by bootstrapping (see methods).

enzymes with >5 visits to Half for RNAP D446A, n=107 enzymes with >5 visits to Half for RNAP wt). As a result of these different kinetics, the RNAP D446A enzyme spends significantly more time in the Pre state throughout pausing compared to wt (Figure 100).

Discussion: These results of RNAP D446A behavior during pausing at the yrbl pause sequence corroborate our overall model and interpretation of the SPRNT RNAP wt pausing results and also help to dissect the importance of the RNAP-G_{cre} interaction for pausing. Although the overall pause lifetime was significantly increased with RNAP D446A (Figure 93), many of the observations during pausing with the two enzymes were consistent. RNAP D446A visited the same five TEC states during pausing (Figure 94), and the dwell times and branching ratios of these states were independent of [GTP] (Figure 97, Table 96). In addition, the dwelltime distribution of the last visit to Post was consistent with GTP incorporation

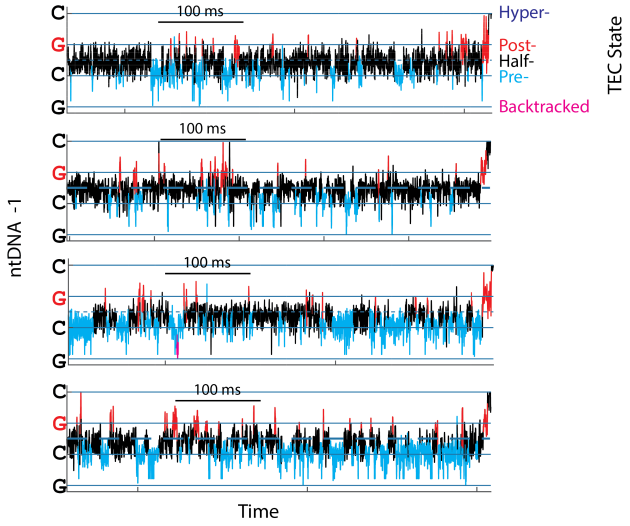


Figure 94: The last 500 ms of four RNAP D446A enzyme traces during yrBL pausing. Data is plotted at 5kHz and acquired with 1000 μM NTPs, 180 mV, 21C. The enzymes visited up to five separate translocation states during pausing, including a Half state in between Pre and Post. The post state was visited prior to pause escape. State transitions were assigned using the point-by-point level finding algorithm outlined in the previous results section..

during this visit. These results support the overall model for pausing as overviewed in the previous results section. The presence of a stable Half state (epTEC) blocks full translocation between Pre and Post, while the very brief duration of the Post state makes GTP binding and incorporation unlikely on any given visit, even at high [GTP].

The data presented here also provides evidence for the presence of the RNAP-G_{CRE} interaction in multiple states throughout pausing. In both the Half and Post states (but not Pre), the dwell time and branching ratio changes between D446A and wt (Figure 100) suggest that D446 hydrogen bonds with guanine at -1 in the ntDNA during both of these states. Losing this residue removes the RNAP-G_{cre} interaction and decreases the stability of both states, reducing the likelihood of transitioning from Half to Post and decreasing the dwell time in Post, which, in combination, reduces the likelihood of pause escape and elongates pause lifetime.

Because the measurements in SPRNT are only sensitive to changes in positioning of the

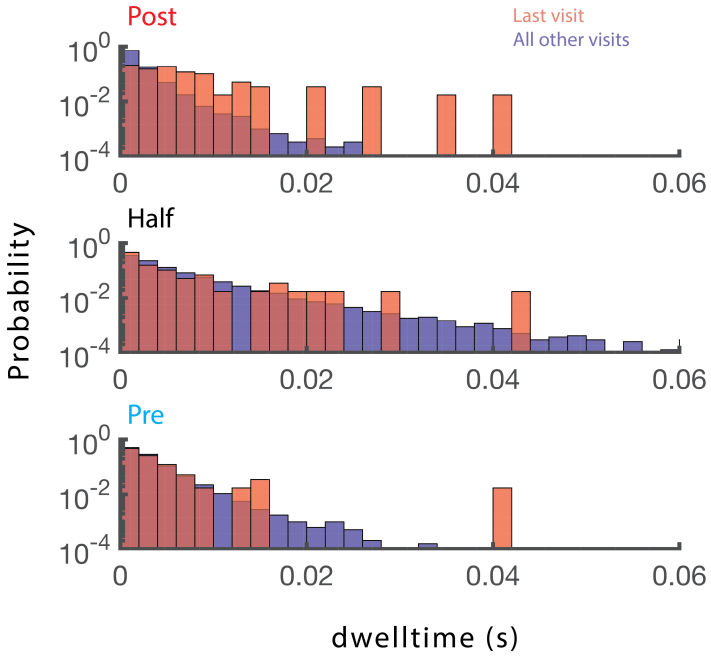


Figure 95: The dwelltime distribution for the last visit to Post prior to pause escape varied from other visits to Post (k.s. test p-value = 3.6e-18), suggesting that reaction follows a different kinetic path during the final visit. The last visit to Post distribution had an increased duration (9.2 /*pm* 1.8 ms vs. 1.9 /*pm* 0.2 ms) and fits the shape of a multi-exponential function, suggesting that there are multiple kinetic steps involved in the transition out of the last visit. The last visits to Pre (k.s. test p-value = 0.97) and Half (k.s. test p-value = 0.46) were indistinguishable from other visits. From this data, we conclude that NTP binding, condensation, and transition into the next Pre state occur during the final visit to Post.

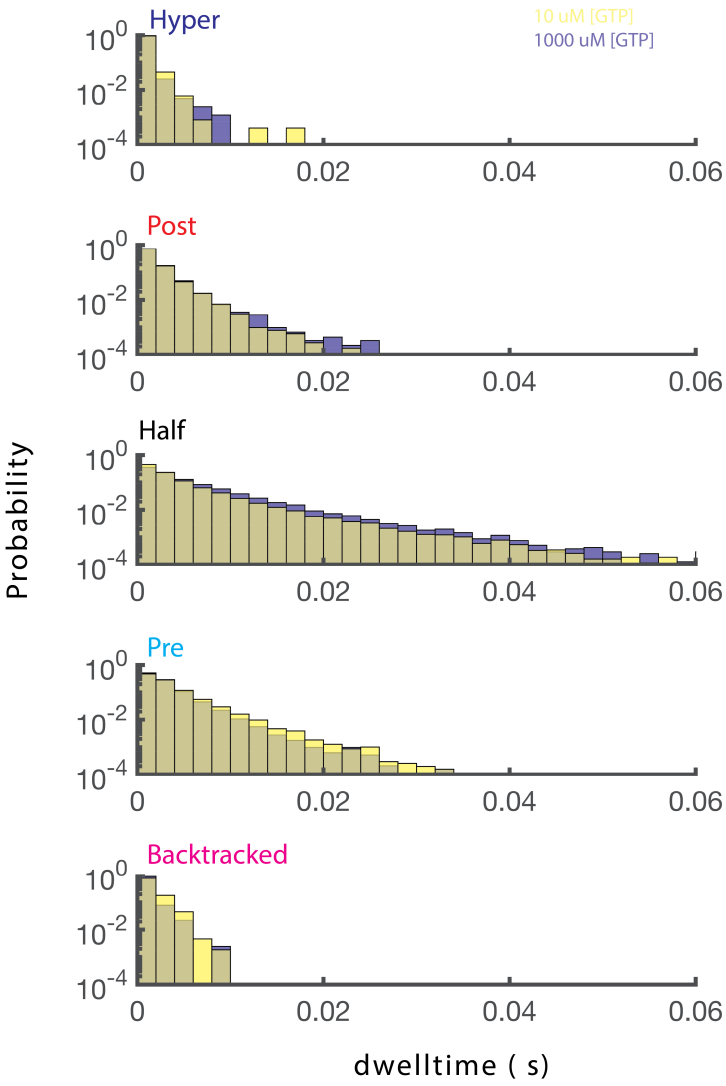


Figure 96: The dwelltime distributions between 10 μM (yellow) and 1000 μM (blue) for all of the TEC states during pausing with RNAP D446A were unchanged. (n=75 enzymes at 1000 μM, n=23 enzymes at 10 μM [GTP]). All data was acquired at 180 mV, 21C.

	10 uM	1000 uM
$\langle t_{Post} \rangle$ =	1.8 ± 0.1 ms	1.9 ± 0.2 ms
$\langle t_{Half} \rangle$ =	4.6 ± 0.6 ms	5.4 ± 0.4 ms
$\langle t_{Pre} \rangle$ =	3.2 ± 0.4 ms	3.2 ± 0.1 ms

	10 uM	1000 uM
$P_{Back, Post}$ =	87.1 ± 3.5 %	91.2 ± 1.9 %
$P_{Back, Half}$ =	65.6 ± 3.8 %	67.1 ± 2.1 %
$P_{Back, Pre}$ =	1.8 ± 0.5 %	2.4 ± 0.4 %

Figure 97: The average dwelltimes and backstep probabilities for the TEC states during pausing with RNAP D446A were not changed significantly between 10 μM (yellow) and 1000 μM (blue). Averages were calculated by finding a mean dwelltime for every enzyme that visited a state at least five times during pausing, creating an average dwell time for each state for each enzyme. The overall average was found by taking a mean and standard deviation of the mean for the population of enzymes in each condition. (n=75 enzymes with at least 5 visits to half at 1000 μM, n=20 enzymes with at least five visits to hal at 10 μM [GTP]).

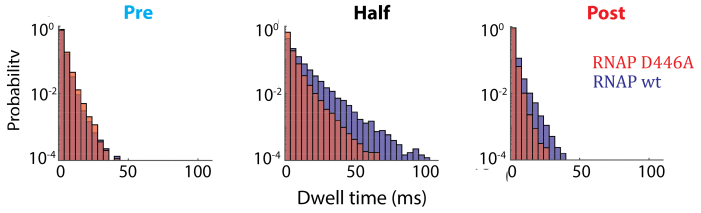


Figure 98: The dwelltime distributions between RNAP D446A (red) and RNAP wt (blue) for the Pre state were unchanged during pausing, while the dwelltime distributions were significantly changed for both half and Post during pausing.We combined the data from different [GTPs] for each enzyme as the dwell times were not affected by [GTP] for both enzymes (n=97 enzymes for RNAP wt, n=65 enzymes for RNAP D446A). All data was acquired at 180 mV, 21C.

	RNAP wt	RNAP D446A
$\langle t_{Post} \rangle$ =	2.2 ± 0.1 ms	1.7 ± 0.1 ms
$\langle t_{Half} \rangle$ =	8.3 ± 0.3 ms	5.2 ± 0.3 ms
$\langle t_{Pre} \rangle$ =	3.2 ± 0.2 ms	3.2 ± 0.1 ms

	RNAP wt	RNAP D446A
$P_{Back, Post}$ =	90.5 ± 0.8 %	89.9 ± 1.7 %
$P_{Back, Half}$ =	49.9 ± 1.8 %	66.7 ± 1.9 %
$P_{Back, Pre}$ =	3.3 ± 0.5 %	2.2 ± 0.3 %

Figure 99: The average dwelltimes for both the Post and Half state decreased with RNAP D446A, while the dwelltime in the Pre state was unchanged. The average probability of transitioning backwards out of the Half state also increases significantly with RNAP D446A. Values were calculated as in Figure 97. We combined the data from different [GTPs] for each enzyme as the dwell times were not affected by [GTP] (n = 65 enzymes with at least five visits to half for RNAP D446A, n=97 enzymes with at least five visits to half for RNAP wt)

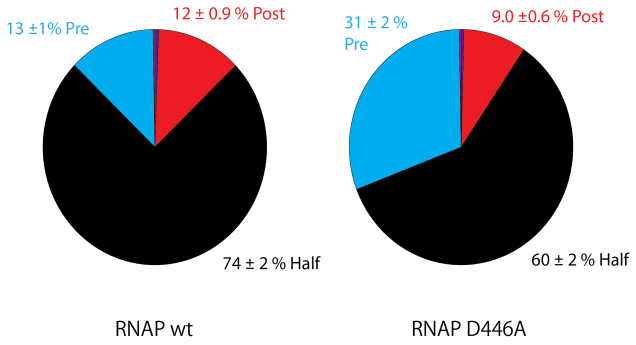


Figure 100: We compared the percent of the total pause lifetime spent in each state between RNAP wt and RNAP D446A. RNAP D446A spends significantly more time in the Pre state compared to wt. The small slice between Pre and Post represents time spent in Back and Hyper and totals less than one percent for both wt and D446A. We combined the data from different [GTPs] for each enzyme as the dwell times were not affected by [GTP] (n = 65 enzymes with at least five visits to half for RNAP D446A, n=97 enzymes with at least five visits to half for RNAP wt)

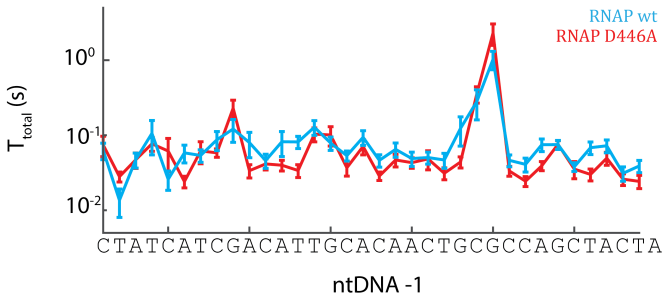


Figure 101: We tracked RNAP D446A outside of pausing, determining the total time spent in each transcription position. (n=29 enzymes). Data was acquired at 1000 μM [NTPs], 180 mV, 21C.

tDNA (not ntDNA) relative to the enzyme, it is difficult to determine from these results whether or not the RNAP-G_{cre} interaction is always present during the Half state, or if the ntDNA guanine could toggle between the -1 and +1 position without corresponding motion of the tilted hybrid and of the upstream tDNA. Further experiments will be needed to determine the exact timing and order of these structural rearrangements during pausing.

It is interesting to note that our data suggests that there are multiple kinetic steps contained within the Half state measured with SPRNT. When the enzyme was in the Half state during pausing, the probability of transitioning to the Pre state was smaller if the previous transition was from Post to Half ($P_{BackgivenBack}$ was 41% while. P_{Back} was 49% for RNAP wt at 180 mV, $P_{BackgivenBack}$ was 61% while P_{Back} was 69% for RNAP D446A at 180 mV). This suggests that there is a hidden kinetic transition during the Half state that does not include repositioning of the tDNA. This could support evidence of ntDNA movement during the Half state, or possibly transitions of other important protein domains that are required for active site rearrangement, like trigger loop folding.

Methods:

The methods for running the RNAP D446A enzyme were identical to those for RNAP core wt.

Errors for the mean pause lifetime were calculated using a bootstrapping method, in which we created 1,000 different pause lifetime distributions by randomly drawing n samples (with replacement) from the original pause lifetime distribution (where, n=number of measurements in initial distribution of pause lifetimes). We calculated a mean lifetime from each of the 1,000 distributions. The error was defined as the standard deviation of these 1,000 mean lifetimes.

0.4.10 RNAP D446A outside of pausing

We also measured the total time spent in each transcription position outside of the pause for RNAP D446A (1000 μM NTPs, 180 mV, 21C, Figure 101). Interestingly, the total times for RNAP D446A were not greater at any of the positions outside of pausing compared to RNAP wt (even at other GTP incorporation positions, where the RNAP-G_{cre} interaction is supposed to play a role). It could be the case that the large assisting force applied in these measurements helped stabilized Post at these positions, so that the probability of GTP incorporation was still very high even without the RNAP-G_{cre} interaction.

0.5 Conclusions

To briefly summarize the information presented in this dissertation, we have developed a new single-molecule technique (SPRNT) that can be used to investigate RNAP transcription at unprecedented spatiotemporal resolution. We have thoroughly outlined the methods and initial results important for employing SPRNT for RNAP studies, enabling high-throughput investigation of RNAP in various experimental conditions (applied force, [NTP]) and transcription behavior at various important DNA sequences.

Using this technique, we monitored many *E. coli* RNAP core complexes during transcription elongation with an assisting force. These experiments revealed single-nucleotide steps of RNAP with millisecond durations for the first time in single-molecule traces. We determined that during transcription elongation at low [NTP] and high assisting force, RNAP primarily stalled in a post-translocated state (Post), with brief deviations forward to a hyper-translocated state (Hyper) and backwards to a pre-translocated state (Pre). The rates and frequencies of these transitions varied significantly with DNA sequence. We performed the first measurements of the underlying rate constants between TEC state transitions at individual incorporation positions by varying the applied force and fitting the resulting changes in rate to our model of transcription elongation. We extended this technique to investigate RNAP pausing at high resolution. During transcription pausing at an elemental pause se-

quence, we observed transitions between five distinct enzyme states (Backtracked, Pre, Half, Post, and Hyper), including a half-translocated state between Pre and Post that had been hypothesized, but could not be directly resolved until the development of SPRNT. We developed a model for RNAP pausing and elongation by varying the applied force and monitoring RNAP mutants with SPRNT.

The results presented in this dissertation demonstrate the capabilities of SPRNT to answer open questions in the field of transcription. With access to individual reaction cycles, we can begin to tease out the subtleties of this complicated system and start to understand how DNA sequence finely controls RNAP behavior. More SPRNT experiments over a wide range of DNA sequences will be required to fully uncover these interactions. Over the coming years, we hope SPRNT will become a common tool in many laboratories to answer these open questions in the field of transcription as well as other types of single-molecule enzymology.

Furthermore, the results gathered from this thesis will also have practical implications for biological engineering. For instance, deeper understanding of transcription pausing and its role in cotranscriptional folding will benefit synthetic biologists. RNA secondary structures are specifically designed by scientists to fold and interact with small molecules. However, proper folding of these structures requires precise understanding of the timing of events in transcription. Perhaps, pausing sequences with fully understood mechanics and timing could be co-opted to improve the implementation of RNA designs. In addition, deeper understanding of the kinetic pathway involved in transcription elongation will improve attempts to manipulate specific reaction steps. In the future, SPRNT may even be used to determine the mechanism in which RNAP interacts with designed transcription drug targets.

0.6 Brief note and acknowledgments

My history in the UW nanopore lab: I joined the UW nanopore lab in the summer of 2011 as an REU student over the summer (part of the Amgen Scholars program). While I only worked with the group for a few months, I enjoyed my experience enough to ask for a full time position in the group after receiving my undergraduate degree from Santa Clara University.

Jens graciously offered me a position; so I joined the lab as a laboratory technician in the fall of 2012, where I ran experiments full time. Over my time as a lab tech, my interest and knowledge in bio engineering continued to grow, as did my responsibilities in the lab. With the guidance of the more experienced researches, I started designing my own experiments and analyzing the data myself. As a natural next step in becoming a scientific researcher, I joined the PhD program in the Molecular Engineering and Sciences Department in the Fall of 2014.

It has been an amazing experience to watch the field of nanopore sequencing grow and prosper, from the work done here at UW and elsewhere in the field of biophysics. In my time in the lab I have been lucky to play a small part of many different projects. From detecting methylations and alternative bases, to helping design artificial membranes for improved stability, to sequencing long reads, to developing SPRNT with studies of hel 308 helicase, I have learned intimately about the field and had the privilege to witness this technology develop first hand.

There are many people to whom I owe thanks for making it this far: My PI Jens Gundlach for teaching me how to be a successful scientist, manager, and to never let ego stand in the way of truth; my mentors, role models, and friends Andrew Laszlo and Ian Derrington who pioneered nanopore sequencing here at UW; my fellow lab mates Jonathan Craig, Henry Brinkerhoff, Matthew Noakes, and Sinduja Marx who are fantastic colleagues, friends, and the future of SPRNT; The other lab techs and researchers, Benjamin Tickman, Kenji Doering, Noah de Leeuw, Hugh Higginbothom, Jonathan Mount, Jasmine Bowman, Katherine Baker, Jesse Huang, Chris Kim, Sarah Abell and others, who truly made these experiments possible; and my collaborators at Rutgers University, Richard Ebright and Abhishek Mazumder, who have enabled this investigation into transcription.

BIBLIOGRAPHY

[1] Carlos Bustamante and Jeffrey R Moffitt. Past, present and future of single-molecule studies of transcription. *RNA Polymerases as Molecular Motors*, pages 302–14, 2009.

[2] Hong Yin, Michelle D Wang, Karel Svoboda, Robert Landick, Steven M Block, and Jeff Gelles. Transcription against an applied force. *Science*, pages 1653–1657, 1995.

[3] David Keller and Carlos Bustamante. The mechanochemistry of molecular motors. *Biophysical Journal*, 78(2):541–556, 2000.

[4] Elio A Abbondanzieri, William J Greenleaf, Joshua W Shaevitz, Robert Landick, and Steven M Block. Direct observation of base-pair stepping by rna polymerase. *Nature*, 438(7067):460, 2005.

[5] Andrew H Laszlo, Ian M Derrington, and Jens H Gundlach. Mspa nanopore as a single-molecule tool: From sequencing to sprnt. *Methods*, 105:75–89, 2016.

[6] Gerald M Cherf, Kate R Lieberman, Hytham Rashid, Christopher E Lam, Kevin Karplus, and Mark Akeson. Automated forward and reverse ratcheting of dna in a nanopore at 5-a precision. *Nature biotechnology*, 30(4):344–348, 2012.

[7] Elizabeth A Manrao, Ian M Derrington, Andrew H Laszlo, Kyle W Langford, Matthew K Hopper, Nathaniel Gillgren, Mikhail Pavlenok, Michael Niederweis, and Jens H Gundlach. Reading dna at single-nucleotide resolution with a mutant mspa nanopore and phi29 dna polymerase. *Nature biotechnology*, 30(4):349–353, 2012.

[8] Ian M Derrington, Jonathan M Craig, Eric Stava, Andrew H Laszlo, Brian C Ross, Henry Brinkerhoff, Ian C Nova, Kenji Doering, Benjamin I Tickman, Mostafa Ronaghi,

et al. Subangstrom single-molecule measurements of motor proteins using a nanopore. *Nature biotechnology*, 33(10):1073–1075, 2015.

[9] Jonathan M Craig, Andrew H Laszlo, Henry Brinkerhoff, Ian M Derrington, Matthew T Noakes, Ian C Nova, Benjamin I Tickman, Kenji Doering, Noah F de Leeuw, and Jens H Gundlach. Revealing dynamics of helicase translocation on single-stranded dna using high-resolution nanopore tweezers. *Proceedings of the National Academy of Sciences*, page 201711282, 2017.

[10] Gongyi Zhang, Elizabeth A Campbell, Leonid Minakhin, Catherine Richter, Konstantin Severinov, and Seth A Darst. Crystal structure of thermus aquaticus core rna polymerase at 3.3 å resolution. *Cell*, 98(6):811–824, 1999.

[11] Bo Huang, Mark Bates, and Xiaowei Zhuang. Super-resolution fluorescence microscopy. *Annual review of biochemistry*, 78:993–1016, 2009.

[12] Henri C Buc and Terence Strick. *RNA polymerases as molecular motors*, volume 16. Royal Society of Chemistry, 2009.

[13] David Dulin, Tao Ju Cui, Jelmer Cnossen, Margreet W Docter, Jan Lipfert, and Nynke H Dekker. High spatiotemporal-resolution magnetic tweezers: Calibration and applications for dna dynamics. *Biophysical journal*, 109(10):2113–2125, 2015.

[14] Seamus J Holden, Stephan Uphoff, Johannes Hohlbein, David Yadin, Ludovic Le Reste, Oliver J Britton, and Achillefs N Kapanidis. Defining the limits of single-molecule fret resolution in tirf microscopy. *Biophysical journal*, 99(9):3102–3111, 2010.

[15] Jeffrey R Moffitt, Yann R Chemla, Steven B Smith, and Carlos Bustamante. Recent advances in optical tweezers. *Annual review of biochemistry*, 77, 2008.

[16] Thorsten Hugel and Markus Seitz. The study of molecular interactions by afm force spectroscopy. *Macromolecular rapid communications*, 22(13):989–1016, 2001.

[17] Karel Svoboda, Christoph F Schmidt, Bruce J Schnapp, and Steven M Block. Direct observation of kinesin stepping by optical trapping interferometry. *Nature*, 365(6448):721–727, 1993.

[18] Jeffrey T Finer, Robert M Simmons, and James A Spudich. Single myosin molecule mechanics: piconewton forces and nanometre steps. *Nature*, 368(6467):113–119, 1994.

[19] Gene-Wei Li and X Sunney Xie. Central dogma at the single-molecule level in living cells. *Nature*, 475(7356):308, 2011.

[20] JD Watson and FHC Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *JAMA*, 269(15):1966–1967, 1993.

[21] John J Kasianowicz, Eric Brandin, Daniel Branton, and David W Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, 93(24):13770–13773, 1996.

[22] Mark Akeson, Daniel Branton, John J Kasianowicz, Eric Brandin, and David W Deamer. Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single rna molecules. *Biophysical journal*, 77(6):3227–3233, 1999.

[23] Tom Z Butler, Mikhail Pavlenok, Ian M Derrington, Michael Niederweis, and Jens H Gundlach. Single-molecule dna detection with an engineered mspa protein nanopore. *Proceedings of the National Academy of Sciences*, 105(52):20647–20652, 2008.

[24] David Stoddart, Andrew J Heron, Ellina Mikhailova, Giovanni Maglia, and Hagan Bayley. Single-nucleotide discrimination in immobilized dna oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences*, 106(19):7702–7707, 2009.

[25] Robert F Purnell, Kunal K Mehta, and Jacob J Schmidt. Nucleotide identification and orientation discrimination of dna homopolymers immobilized in a protein nanopore. *Biophysical Journal*, 96(3):649a, 2009.

[26] Elizabeth A Manrao, Ian M Derrington, Mikhail Pavlenok, Michael Niederweis, and Jens H Gundlach. Nucleotide discrimination with dna immobilized in the mspa nanopore. *PloS one*, 6(10):e25723, 2011.

[27] Kate R Lieberman, Gerald M Cherf, Michael J Doody, Felix Olasagasti, Yvette Kolodji, and Mark Akeson. Processive replication of single dna molecules in a nanopore catalyzed by phi29 dna polymerase. *Journal of the American Chemical Society*, 132(50):17961, 2010.

[28] Ian C Nova, Ian M Derrington, Jonathan M Craig, Matthew T Noakes, Benjamin I Tickman, Kenji Doering, Hugh Higinbotham, Andrew H Laszlo, and Jens H Gundlach. Investigating asymmetric salt profiles for nanopore dna sequencing with biological porin mspa. *PloS one*, 12(7):e0181599, 2017.

[29] Ian M Derrington, Tom Z Butler, Marcus D Collins, Elizabeth Manrao, Mikhail Pavlenok, Michael Niederweis, and Jens H Gundlach. Nanopore dna sequencing with mspa. *Proceedings of the National Academy of Sciences*, 107(37):16060–16065, 2010.

[30] Andrew H Laszlo, Ian M Derrington, Henry Brinkerhoff, Kyle W Langford, Ian C Nova, Jenny Mae Samson, Joshua J Bartlett, Mikhail Pavlenok, and Jens H Gundlach. Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore mspa. *Proceedings of the National Academy of Sciences*, 110(47):18904–18909, 2013.

[31] Jonathan M Craig, Andrew H Laszlo, Ian M Derrington, Brian C Ross, Henry Brinkerhoff, Ian C Nova, Kenji Doering, Benjamin I Tickman, Mark T Svet, and Jens H Gundlach. Direct detection of unnatural dna nucleotides dnam and d5sics using the mspa nanopore. *PloS one*, 10(11):e0143253, 2015.

[32] Andrew H Laszlo, Ian M Derrington, Brian C Ross, Henry Brinkerhoff, Andrew Adey, Ian C Nova, Jonathan M Craig, Kyle W Langford, Jenny Mae Samson, Riza Daza, et al. Decoding long nanopore sequencing reads of natural dna. *Nature biotechnology*, 32(8):829–833, 2014.

[33] Philip M Ashton, Satheesh Nair, Tim Dallman, Salvatore Rubino, Wolfgang Rabsch, Solomon Mwaigwisya, John Wain, and Justin O'grady. Minion nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature biotechnology*, 33(3):296–300, 2015.

[34] Joshua Quick, Nicholas J Loman, Sophie Duraffour, Jared T Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, et al. Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228, 2016.

[35] Kate R Lieberman, Joseph M Dahl, Ai H Mai, Mark Akeson, and Hongyun Wang. Dynamics of the translocation step measured in individual dna polymerase complexes. *Journal of the American Chemical Society*, 134(45):18816, 2012.

[36] Kate R Lieberman, Joseph M Dahl, Ai H Mai, Ashley Cox, Mark Akeson, and Hongyun Wang. Kinetic mechanism of translocation and dntp binding in individual dna polymerase complexes. *Journal of the American Chemical Society*, 135(24):9149, 2013.

[37] Nataliya Korzheva, Arkady Mustaev, Maxim Kozlov, Arun Malhotra, Vadim Nikiforov, Alex Goldfarb, and Seth A Darst. A structural model of transcription elongation. *Science*, 289(5479):619–625, 2000.

[38] P Cramer. Common structural features of nucleic acid polymerases. *Bioessays*, 24(8):724–729, 2002.

[39] Irina O Vvedenskaya, Hanif Vahedian-Movahed, Jeremy G Bird, Jared G Knoblauch, Seth R Goldman, Yu Zhang, Richard H Ebright, and Bryce E Nickels. Interactions

between rna polymerase and the core recognition element counteract pausing. *Science*, 344(6189):1285–1289, 2014.

[40] Evgeny Nudler, Arkady Mustaev, Alex Goldfarb, and Evgeny Lukhtanov. The rna–dna hybrid maintains the register of transcription by preventing backtracking of rna polymerase. *Cell*, 89(1):33–41, 1997.

[41] Natalia Komissarova and Mikhail Kashlev. Transcriptional arrest: Escherichia coli rna polymerase translocates backward, leaving the 3 end of the rna intact and extruded. *Proceedings of the National Academy of Sciences*, 94(5):1755–1760, 1997.

[42] Michelle D Wang, Mark J Schnitzer, Hong Yin, Robert Landick, Jeff Gelles, and Steven M Block. Force and velocity measured for single molecules of rna polymerase. *Science*, 282(5390):902–907, 1998.

[43] R John Davenport, Gijs JL Wuite, Robert Landick, and Carlos Bustamante. Single-molecule study of transcriptional pausing and arrest by e. coli rna polymerase. *Science*, 287(5462):2497, 2000.

[44] Keir C Neuman, Elio A Abbondanzieri, Robert Landick, Jeff Gelles, and Steven M Block. Ubiquitous transcriptional pausing is independent of rna polymerase backtracking. *Cell*, 115(4):437–447, 2003.

[45] Matthew H Larson, Jing Zhou, Craig D Kaplan, Murali Palangat, Roger D Kornberg, Robert Landick, and Steven M Block. Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of rna polymerase ii. *Proceedings of the National Academy of Sciences*, 109(17):6555–6560, 2012.

[46] Lu Bai, Robert M Fulbright, and Michelle D Wang. Mechanochemical kinetics of transcription elongation. *Physical review letters*, 98(6):068103, 2007.

[47] Y Whitney Yin and Thomas A Steitz. The structural mechanism of translocation and helicase activity in t7 rna polymerase. *Cell*, 116(3):393–404, 2004.

[48] Richard Guajardo and Rui Sousa. A model for the mechanism of polymerase translocation. *Journal of molecular biology*, 265(1):8–19, 1997.

[49] Natalia Komissarova and Mikhail Kashlev. Rna polymerase switches between inactivated and activated states by translocating back and forth along the dna and the rna. *Journal of Biological Chemistry*, 272(24):15329–15338, 1997.

[50] Lu Bai, Alla Shundrovsky, and Michelle D Wang. Sequence-dependent kinetic model for transcription elongation by rna polymerase. *Journal of molecular biology*, 344(2):335–349, 2004.

[51] Gil Bar-Nahum, Vitaly Epshtein, Andrei E Ruckenstein, Ruslan Rafikov, Arkady Mustaev, and Evgeny Nudler. A ratchet mechanism of transcription elongation and its control. *Cell*, 120(2):183–193, 2005.

[52] P Cramer, K-J Armache, S Baumli, S Benkert, F Brueckner, C Buchen, GE Damsma, S Dengl, SR Geiger, AJ Jasiak, et al. Structure of eukaryotic rna polymerases. *Annu. Rev. Biophys.*, 37:337–352, 2008.

[53] Vasisht R Tadigotla, Dáibhid Ó Maoiléidigh, Anirvan M Sengupta, Vitaly Epshtein, Richard H Ebright, Evgeny Nudler, and Andrei E Ruckenstein. Thermodynamic and kinetic modeling of transcriptional pausing. *Proceedings of the National Academy of Sciences of the United States of America*, 103(12):4439–4444, 2006.

[54] Manchuta Dangkulwanich, Toyotaka Ishibashi, Shixin Liu, Maria L Kireeva, Lucyna Lubkowska, Mikhail Kashlev, and Carlos J Bustamante. Complete dissection of transcription elongation reveals slow translocation of rna polymerase ii in a linear ratchet mechanism. *Elife*, 2:e00971, 2013.

[55] Yara X Mejia, Evgeny Nudler, and Carlos Bustamante. Trigger loop folding determines transcription rate of escherichia colis rna polymerase. *Proceedings of the National Academy of Sciences*, 112(3):743–748, 2015.

[56] Robert Landick, Jannette Carey, and Charles Yanofsky. Translation activates the paused transcription complex and restores transcription of the trp operon leader region. *Proceedings of the National Academy of Sciences*, 82(14):4663–4667, 1985.

[57] Tao Pan and Tobin Sosnick. Rna folding during transcription. *Annu. Rev. Biophys. Biomol. Struct.*, 35:161–175, 2006.

[58] J Kenneth Wickiser, Wade C Winkler, Ronald R Breaker, and Donald M Crothers. The speed of rna transcription and metabolite binding kinetics operate an fmn riboswitch. *Molecular cell*, 18(1):49–60, 2005.

[59] GA Kassavetis and MJ Chamberlin. Pausing and termination of transcription within the early region of bacteriophage t7 dna in vitro. *Journal of Biological Chemistry*, 256(6):2777–2786, 1981.

[60] Ivan Gusarov and Evgeny Nudler. The mechanism of intrinsic transcription termination. *Molecular cell*, 3(4):495–504, 1999.

[61] John P Richardson. Rho-dependent termination and atpases in transcript termination. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, 1577(2):251–260, 2002.

[62] Qiang Zhou, Tiandao Li, and David H Price. Rna polymerase ii elongation control. *Annual review of biochemistry*, 81:119–143, 2012.

[63] Ravindra V Dalal, Matthew H Larson, Keir C Neuman, Jeff Gelles, Robert Landick, and Steven M Block. Pulling on the nascent rna during transcription does not alter kinetics of elongation or ubiquitous pausing. *Molecular cell*, 23(2):231–239, 2006.

[64] Kristina M Herbert, Arthur La Porta, Becky J Wong, Rachel A Mooney, Keir C Neuman, Robert Landick, and Steven M Block. Sequence-resolved detection of pausing by single rna polymerase molecules. *Cell*, 125(6):1083–1094, 2006.

[65] Eric A Galburt, Stephan W Grill, Anna Wiedmann, Lucyna Lubkowska, Jason Choy, Eva Nogales, Mikhail Kashlev, and Carlos Bustamante. Backtracking determines the force sensitivity of rnap ii in a factor-dependent manner. *Nature*, 446(7137):820–823, 2007.

[66] Yara X Mejia, Hanbin Mao, Nancy R Forde, and Carlos Bustamante. Thermal probing of e. coli rna polymerase off-pathway mechanisms. *Journal of molecular biology*, 382(3):628–637, 2008.

[67] Martin Depken, Eric A Galburt, and Stephan W Grill. The origin of short transcriptional pauses. *Biophysical journal*, 96(6):2189–2193, 2009.

[68] Courtney Hodges, Lacramioara Bintu, Lucyna Lubkowska, Mikhail Kashlev, and Carlos Bustamante. Nucleosomal fluctuations govern the transcription dynamics of rna polymerase ii. *Science*, 325(5940):626–628, 2009.

[69] Maria L Kireeva and Mikhail Kashlev. Mechanism of sequence-specific pausing of bacterial rna polymerase. *Proceedings of the National Academy of Sciences*, 106(22):8900–8905, 2009.

[70] R Landick. The regulatory roles and mechanism of transcriptional pausing, 2006.

[71] Robert Landick. Transcriptional pausing without backtracking. *Proceedings of the National Academy of Sciences*, 106(22):8797–8798, 2009.

[72] Jing Zhou, Kook Sun Ha, Arthur La Porta, Robert Landick, and Steven M Block. Applied force provides insight into transcriptional pausing and its modulation by transcription factor nusa. *Molecular cell*, 44(4):635–646, 2011.

[73] Albert Weixlbaumer, Katherine Leon, Robert Landick, and Seth A Darst. Structural basis of transcriptional pausing in bacteria. *Cell*, 152(3):431–441, 2013.

[74] Matthew H Larson, Rachel A Mooney, Jason M Peters, Tricia Windgassen, Dhananjaya Nayak, Carol A Gross, Steven M Block, William J Greenleaf, Robert Landick, and Jonathan S Weissman. A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science*, 344(6187):1042–1047, 2014.

[75] L Stirling Churchman and Jonathan S Weissman. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330), 2011.

[76] Irina O Vvedenskaya, Hanif Vahedian-Movahed, Yuanchao Zhang, Deanne M Taylor, Richard H Ebright, and Bryce E Nickels. Interactions between rna polymerase and the core recognition element are a determinant of transcription start site selection. *Proceedings of the National Academy of Sciences*, 113(21):E2899–E2905, 2016.

[77] Yuhong Zuo and Thomas A Steitz. Crystal structures of the e. coli transcription initiation complexes with a complete bubble. *Molecular cell*, 58(3):534–540, 2015.

[78] Jin Young Kang, Paul Dominic B Olinares, James Chen, Elizabeth A Campbell, Arkady Mustaev, Brian T Chait, Max E Gottesman, and Seth A Darst. Structural basis of transcription arrest by coliphage hk022 nun in an escherichia coli rna polymerase elongation complex. *Elife*, 6:e25478, 2017.

[79] Xieyang Guo, Alexander G Myasnikov, James Chen, Corinne Crucifix, Gabor Papai, Maria Takacs, Patrick Schultz, and Albert Weixlbaumer. Structural basis for nusa stabilized transcriptional pausing. *Molecular cell*, 69(5):816–827, 2018.

[80] Jin Young Kang, Tatiana V Mishanina, Michael J Bellecourt, Rachel Anne Mooney, Seth A Darst, and Robert Landick. Rna polymerase accommodates a pause rna hairpin by global conformational rearrangements that prolong pausing. *Molecular cell*, 69(5):802–815, 2018.

[81] Alan CM Cheung, Sarah Sainsbury, and Patrick Cramer. Structural basis of initial rna polymerase ii transcription. *The EMBO journal*, 30(23):4755–4763, 2011.

[82] Bo Shu and Peng Gong. Structural basis of viral rna-dependent rna polymerase catalysis and translocation. *Proceedings of the National Academy of Sciences*, 113(28):E4005–E4014, 2016.

[83] Yu Zhang, Yu Feng, Sujoy Chatterjee, Steve Tuske, Mary X Ho, Eddy Arnold, and Richard H Ebright. Structural basis of transcription initiation. *Science*, 338(6110):1076–1080, 2012.

[84] Evgeny Nudler and Max E Gottesman. Transcription termination and anti-termination in e. coli. *Genes to Cells*, 7(8):755–768, 2002.

[85] Elena A Lesnik, Rangarajan Sampath, Harold B Levene, Timothy J Henderson, John A McNeil, and David J Ecker. Prediction of rho-independent transcriptional terminators in escherichia coli. *Nucleic Acids Research*, 29(17):3583–3594, 2001.

[86] Jennifer C McDowell, Jeffrey W Roberts, Ding Jun Jin, Carol Gross, et al. Determination of intrinsic transcription termination efficiency by rna polymerase elongation rate. *SCIENCE-NEW YORK THEN WASHINGTON-*, pages 822–822, 1994.

[87] Natalia Komissarova, Jodi Becker, Stephanie Solter, Maria Kireeva, and Mikhail Kashlev. Shortening of rna: Dna hybrid in the elongation complex of rna polymerase is a prerequisite for transcription termination. *Molecular cell*, 10(5):1151–1162, 2002.

[88] Thomas J Santangelo and Jeffrey W Roberts. Forward translocation is the natural pathway of rna release at an intrinsic terminator. *Molecular cell*, 14(1):117–126, 2004.

[89] Matthew H Larson, William J Greenleaf, Robert Landick, and Steven M Block. Applied force reveals mechanistic and energetic details of transcription termination. *Cell*, 132(6):971–982, 2008.

[90] Sinan Arslan, Rustem Khafizov, Christopher D Thomas, Yann R Chemla, and Taekjip Ha. Engineering of a superhelicase through conformational control. *Science*, 348(6232):344–347, 2015.

[91] Francine Toulmé, Christine Mosrin-Huaman, Jason Sparkowski, Asis Das, Marc Leng, and A Rachid Rahmouni. Grea and greb proteins revive backtracked rna polymerase in vivo by promoting transcript trimming. *The EMBO journal*, 19(24):6853–6859, 2000.

[92] Jeffrey Roberts and Joo-Seop Park. Mfd, the bacterial transcription repair coupling factor: translocation, repair and termination. *Current opinion in microbiology*, 7(2):120–125, 2004.

[93] Matthew T Noakes, Henry Brinkerhoff, Andrew H Laszlo, Ian M Derrington, Kyle W Langford, Jonathan W Mount, Jasmine L Bowman, Katherine S Baker, Kenji M Doering, Benjamin I Tickman, et al. Increasing the accuracy of nanopore dna sequencing using a time-varying cross membrane voltage. *Nature biotechnology*, 37(6):651–656, 2019.

[94] Lalit Bahl, John Cocke, Frederick Jelinek, and Josef Raviv. Optimal decoding of linear codes for minimizing symbol error rate (corresp.). *IEEE Transactions on information theory*, 20(2):284–287, 1974.

[95] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.