Title
Improving the resolution and accuracy of optical tweezers through algorithmic and instrumental advances

Author
Lee, Antony Ann-Tzer

Publication Date
2018

Peer reviewed|Thesis/dissertation

Improving the resolution and accuracy of optical tweezers through algorithmic and instrumental advances

By

Antony Ann-Tzer Lee

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Physics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Carlos Bustamante, Chair
Professor Xavier Darzacq
Professor Hernan Garcia-Melan

Spring 2018

Improving the resolution and accuracy of optical tweezers
through algorithmic and instrumental advances

Abstract

Improving the resolution and accuracy of optical tweezers
through algorithmic and instrumental advances

by

Antony Ann-Tzer Lee

Doctor of Philosophy in Physics

University of California, Berkeley

Professor Carlos Bustamante, Chair

In the first half of this thesis, we describe our study of the elongation dynamics of *E. coli* RNA polymerase using optical tweezers. Optical tweezers constitute an important tool in modern biophysical research, as they allow the manipulation and tracking of individual molecules, such as enzymes that carry out diverse biological functions by converting chemical energy into mechanical work. Improvements to the spatio-temporal resolution and accuracy of optical tweezers therefore directly impact our ability to probe the tiniest and fastest motions of such enzymes.

RNA polymerase is a central enzyme present in all organisms, that transcribes the genetic information encoded in DNA into RNA, one nucleotide at a time. This process constitutes the first step of gene expression, and is highly regulated at all its stages: initiation, elongation, and termination. In particular, elongation—i.e., the processive polymerization of the nascent RNA chain—does not occur in a continuous fashion, but consists of periods of active translocation interspersed by long-lived, sequence-dependent pauses, that have been implicated in various biological roles.

While optical tweezers have long been able to observe such long-lived pausing events, many questions remained open, due to the limited spatio-temporal resolution of the technique. Here, we demonstrate algorithmic and instrumental developments that improve our ability to probe the transcription cycle at the finest level. Improvements in spatial resolution allowed us to robustly observe individual translocation events over long distances, and thus record the distribution of the dwell times spent at each position by the enzyme. Improvements in temporal resolution and spatial accuracy allowed us to understand the dynamics of the enzymes immediately as it reaches a "pause site". Specifically, we were able to show that transcription through a pause site is always accompanied by a decrease of the forward transcription rate. We established that entry into "backtracked" pauses occurs in a stepwise fashion, with a relatively slow entry into deeply backtracked states. We also probed the effect

of nascent RNA structures on RNAP dynamics, and found that, depending on the sequence context, such structures could either enhance or attenuate pre-existing pauses.

In the second half of this thesis, we review another fundamental single-molecule technique: super-resolution microscopy. Unlike optical tweezers, optical microscopy allows us to observe cellular processes in vivo or in situ; and the recent development of super-resolution microscopy has greatly enhanced the field of application of the technique. However, super-resolution microscopy also yields data that is much more difficult to interpret than classical ("diffraction-limited") microscopy. We discuss recent developments in our ability, not only to localize molecules with high accuracy, but also to quantify them. Finally, we present a fluorescent protein engineering work, regarding the development of a split-photoactivatable fluorescent protein system, towards the goal of studying protein-protein interaction at high resolution.

This dissertation is dedicated to all my friends,
without whom none of this work would have been possible.

Acknowledgments

First and foremost, my thanks goes to my advisor, Carlos Bustamante, for his continuous support and the freedom that he gave me to pursue diverse projects in his lab. Through my interactions with him, I became an independent scientist, learnt how to push forward a research project, and how to defend my scientific choices. I will look back at my time in his lab as an invaluable stage in my professional development.

Next, I would like to thank the members of my qualifying examination and thesis committee, Xavier Darzacq, Hernan Garcia, and Mike deWeese. They have all offered me extremely important advice, sometimes on short notice, regarding both the pursuit of my graduate research and my future professional choices. I am grateful for the insights that they have offered me, as they will shape my coming years for the better.

A special thought goes to Ignacio Tinoco. Although I only tangentially worked with him, his down-to-earth approach to science and healthy skepticism will always remind me that the scientific endeavour must first and foremost remain anchored in reality. He will be sorely missed.

Carlos takes great pride at assembling a diverse team of talented scientists in his lab, and it is though my interactions with them that I picked up not only many practical lab skills, but also, more importantly, the ability to think like a biophysicist. During the first half of my Ph. D., I learnt super-resolution microscopy from Sang-Hyuk and Jae-Yen and molecular biology from Alyssa and Young-Woo, and started practicing mentoring skills by working with two great undergrads, Émilie and Pengning. Later, as I came back towards optical tweezers, encouraged by Troy and Luis, my main professional colleague was of course Ronen, who deserves a special place in this acknowledgement. Without his help, I would not have been able to reach the point I am at today. Not only he is truly an amazing scientist and mentor, and the breadth of his skill set is absolutely unmatched; but he is also a great friend, and I will remember with fondness our expeditions to Half Dome and to Telescope Peak. Cristhian played an essential role in bringing the base-pair resolution project to its completion; that manuscript would never have reached its end without his hard work! He is of course also a friendly roommate, whose company I enjoy. Although Tingting only briefly worked with us, her advice, whether in the scientific domain or in the professional one, was very helpful. Enze is my spiritual successor in the lab, and I wish him the best! In the analysis

side, he will have a great colleague in Alex. The transcription team included many other great friends and colleagues—Pim, who always fulfilled her chores as "lab queen" in a most helpful manner; Yves and Toyotaka, and their dry sense of humor; César, who shared my experience of moving from microscopy to transcription; Zhijie—we should plan our next eclipse trip together; Omar, Robert, and Alan, who bring a youthful and joyful new touch to the lab.

I have interacted in an on-and-off (two-state?) manner with the other projects in the lab, especially translation. But Dan, Varsha, Wee, and Lisa were much more than colleagues, they were good friends, who always listened sympathetically to my complaints and offered helpful advice. Their contagious happiness always put the lab in a good mood. I also thank Wee for his careful proofreading of this manuscript. I have always appreciated Yumeng's sense of humor. There is no way anyone in the lab can thank Shannon for everything she are doing for all of us. In particular, I will remember the very large amount of time she spent to help me prepare my qualifying exam on a very short notice. I had many long discussions about the meaning of life and of graduate studies (or in the other order) with Bibiana; I will miss her! I am looking forward to see Sara again, perhaps in Amsterdam, although there we may replace the sake tasting with some other beverages. Of course, we may also get to meet Maya, Sam, and Hendrik there, for a b-lab get-together.

Although the older generation has long left the lab and moved towards their next professional stage, I do not forget their advice when I joined the lab. In particular, I want to thank Craig, Gheorghe, and Shixin, who were my very first colleagues here.

It was my honor to be invited by Steve Pressé to work with him on a review of super-resolution data analysis methods. This review greatly widened my knowledge of the technique, and offered me additional insights on a theorist's view of biophysical problems. Moreover, Steve also offered me excellent advice regarding my future career choices, for which I am grateful.

I was very lucky to have the chance to collaborate with Dave Savage's lab. Dave has welcomed me to his lab as one of his own students, and was always helpful. Avi is a great friend and I will miss our scientific and non-scientific discussions; it is too bad that we were not able to finish our joint work.

Outside of the lab, I would like thank in particular my bridge-playing friends, with whom I have shared many of my days when hopelessly trying to think about something else than single molecules: Debbie, for making me a much better player; all my partners—Jannes, Martin, Ryan, Rebecca, Roger, and the others—who had to endure my bids and card play.

Finally, I would like to thank my parents for their continuous support—you know that I always think of you, and am looking forward to come back in France—my siblings, Dominique and Caroline, for their cheerful camaraderie, and of course, Sonia, to whom I remain forever thankful for teaching me that some events, even though *improbable*, may be successfully observed in a single experiment.

Contents

List of figures

Introduction

> Then a man named Tycho Brahe evolved a way of answering the question [of whether the planets went around the sun]. He thought that it might perhaps be a good idea to look very very carefully and to record exactly were the planets appear in the sky, and then the alternative theories might be distinguished from one another.
>
> Richard Feynman,
> *The Character of Physical Law*

Observation is a fundamental part of the scientific inquiry: the precision of the measurements that we perform ultimately decides whether we can distinguish a correct theory from an incorrect one. In this thesis, we describe our work in improving the resolution and accuracy of a single-molecule biophysical technique, optical tweezers, through algorithmic and instrumental developments. We also review recent algorithmic advances in another single-molecule technique, super-resolution microscopy, likewise aimed at furthering the resolution of the method; finally, we present a fluorescent protein engineering work that aimed at providing a new modality to probe protein-protein interactions with high spatial resolution and accuracy.

RNA polymerase and its dynamics

The biological object of our optical tweezers study is the *E. coli* RNA polymerase during its elongation phase. Because the precision that we need to achieve is fundamentally set by the dynamical properties of the object of our study, we first briefly review here the fundamental properties of this enzyme.

RNA polymerase (RNAP) is an essential enzyme present in all living organisms, that performs the first step of the central dogma of molecular biology, namely the transcription of genetic information encoded in DNA into RNA (figure 1) [1]. Due to the central role that the enzyme plays, all three stages of its activity—initiation, elongation, and termination—are tightly regulated.

Briefly, RNAP first binds to specific sites in the genome, known as promoters, to initiate transcription. In bacteria, σ factors of varying promoter specificities [2] transiently associate with the five-subunit ($\alpha_2\beta\beta'\omega$) core enzyme [3] in order to recognize the promoter. RNAP binds the DNA template initially in the closed promoter form (RP_c), melts the DNA to form a transcription bubble, isomerizes into the open form (RP_o) [4], and starts polymerizing an

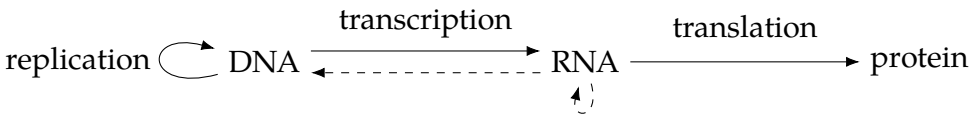


Figure 1:
The central dogma of molecular biology: Genetic information encoded in DNA is replicated, and is transcribed into RNA, which is itself translated into proteins. Other ("special") transfers, marked by dashed arrows—reverse transcription of RNA into DNA, and RNA replication—also occur.

RNA chain. For multiple cycles, this polymerization is abortive, i.e., stops with the release of a short product less than 10 bp in length [5]; however, RNAP ultimately clears the promoter and transitions into the processive elongation phase.

During the elongation phase, RNAP repeatedly alternates between performing a mechanical activity—translocation along the DNA template by a single base pair ($\sim 0.34\,\mathrm{nm}$)—and a chemical activity—condensation of a single NTP into the growing RNA chain. While this process normally occurs at a velocity of a few dozen bp/s, it is also interspersed with long-lived [6–8], sequence-dependent [9–12] pauses, which may last up to dozens of seconds [13]. Throughout the years, these pauses have been implicated in various biological functions, such as regulation of operon expression [14], RNA folding and processing [15], coupling of transcription with translation [16], and termination [16]. Proteins such as NusA [17], NusG [18], GreA, and GreB [19] modulate various pausing pathways. The processes involved in pausing have been widely studied using structural [20] and dynamical [21] methods.

The last stage of the transcription cycle, termination, occurs when RNAP reaches a termination sequence, and consists of the disengagement of RNAP from the DNA template and the release of the nascent RNA. Although termination corresponds to the end of a transcription cycle, it is also highly regulated; for example, premature termination allows for the downregulation of certain operons [22]. In certain cases, termination is catalyzed by the Rho factor, which is itself regulated through interactions with factors such as NusA and NusG [23]. In other cases, termination occurs through the interaction with hairpins in the nascent RNA [24]; there, regulation may occur, for example, via the binding of anti-termination proteins that prevent the formation of terminator hairpins [25].

Resolution and accuracy of optical tweezers

Optical trapping is a phenomenon whereby the radiation pressure of a focused laser beam generates a force gradient that can stabilize micron-sized dielectric particles at a given position. In the simplest approximation, for a dielectric particle of polarizability α, an electric field $\mathbf{E}$

(such as the one present in the focused beam) will induce a dipole moment $\mathbf{p} = \alpha\mathbf{E}$; thus, such a particle will be subjected to a potential energy landscape of the form $-\mathbf{p} \cdot \mathbf{E} = -\alpha\mathbf{E}^2$. The particle will thus be driven towards regions of high electric field, i.e., the beam focus.

Optical trapping was first developed as a tool for the study of the physical properties of radiation pressure [26, 27]. For fundamental physicists, it has over the years evolved into a technique to "cool and trap atoms with laser light", for which Steven Chu, Claude Cohen-Tannoudji, and William D. Phillips were awarded the Nobel Prize in Physics in 1997. But this tool also found an important application in biology, as it allowed the manipulation of individual microorganisms [28] and then of single biological molecules [29]. In particular, optical traps (or "optical tweezers") could be used to track the motion of *molecular motors*, an important class of enzymes that convert chemical energy into mechanical work and displacement [30]. As mentioned above, RNA polymerase is an example of such an enzyme, and it has successfully been studied using optical tweezers [21, 31]. Experiments typically consist in tethering the enzyme to a bead held in an optical trap, in a geometry such that the enzyme's motion is transduced into a measurable displacement of the bead relative to the trap. Other molecular motors that have likewise been "tweezed on" include DNA polymerases [32], DNA translocases [33], the ribosome [34], proteases [35], dynein [36] and kinesin [37], etc. Moreover, the ability of optical tweezers to apply force on the enzyme, and thus to modulate the energy landscape associated with the enzyme's displacement, has yielded many additional insights into the way these molecules perform their mechanochemical coupling.

In order to follow the activity of such enzymes, it is of utmost importance that the instrument possesses both high spatial and temporal resolution (i.e., is able to distinguish physically small or temporally close events), but also high spatial accuracy (i.e., is able to report the exact position at which a certain event occurred). Here, we separately discuss the relevance of each of these requirements, and the approaches we undertook to fulfill them.

Spatial resolution

From an instrumental point of view, an important factor to maximize accuracy and resolution is the *stability* of the instrument, that is, the lack of apparent changes in signal (i.e., *baseline drift*) if the underlying enzyme does not, indeed, move. In the classical, "single-trap" optical tweezers design, a major source of drift comes from the relative, mechanical motion of the optical trap (holding for example the enzyme) relative to the microfluidic chamber where the experiment is performed (and that serves as reference frame for the enzyme's substrate) [38]. An important early advance in stabilizing optical tweezers was the development of the *differential detection* strategy [38], whereby the enzyme and its substrate are each tethered to their own bead held in two optical traps formed by the same laser. In such a design,

mechanical motion of the optical train equally affects both traps, and is cancelled out. In exceptional cases, this design may be stable enough to occasionally observe stepping of an RNA polymerase artificially slowed by NTP deprivation [39]; however, such observations remain exceptionally rare.

In the most commonly used differential detection design (the *differential path* design), the two traps were formed using the two polarization components of the laser beam, by momentarily separating these components with a polarizing beam splitter, reflecting them on two independent piezo-actuated mirrors for trap steering, and recombining them with a second polarizing beam splitter. In chapter I.3, we establish that the relative drift between the two mirrors is the next major source of drift in the system, and we find in chapter I.4 that a different design, based on the fast switching of a single trap between two different positions (the *time-shared* design) [40], does not suffer from this drift. Thus, the use of a time-shared instrument greatly improves the quality of our data by decreasing its baseline drift.

The next step in resolving single base pair motion lies in algorithmic developments. So far, many studies of molecular motors on optical tweezers that attempted to resolve individual steps taken by the motor aimed at understanding basic properties of the enzyme, such as its previously unknown step size (11 bp for the NS3 helicase [41]; 10 bp for the phage φ29 packaging motor [42]; 1 nm for the ClpXP protease [43]). Other studies, on motors with a known step size (1 bp for RNA polymerase [39]; 1 codon for the ribosome [34]) did not adapt their analysis routines to take advantage of that knowledge. It is intuitive that incorporation of a known step size can help extract more information from the experimental data; in general, such incorporation occurs via the formulation of an explicit *generative model*, which lists all sources of signal and of randomness in the data [44].

In chapter I.5, we introduce a generative model for transcription elongation data—in fact, more generally, for motors taking steps of a single size—and show how it can be fitted using the so-called large state-space hidden Markov model [45]. In chapter I.6, we test the resolution of our approach by generating simulated stepping data directly using the optical traps, and show that our approach can indeed accurately recover single stepping events with a size of 1 bp. Finally, in chapter I.7, we apply the framework we developed to actual (slowed) transcription elongation data, and show that we are again able to recover single stepping events. Ultimately, this capability may be used to further our understanding of the dynamics surrounding the nucleotide addition cycle, as well as how this cycle is affected by various transcription inhibitors, which form an important class of drugs [46]. As a proof of concept, we also performed our experiments in the presence of the transcription inhibitor pyrophosphate, and measured how it affected the distribution of dwell times between consecutive steps.

Spatial accuracy

As noted earlier, the elongation dynamics of RNA polymerase, and in particular its pausing, depend on the sequence of the underlying DNA template. In order to study this dependence, it is necessary to develop highly *accurate* measurements, i.e., where the absolute position of the enzyme along its substrate is well characterized, even if at the cost of not being able to consistently resolve single steps. Achieving this goal requires a priori a precise calibration ($\lesssim 0.1\,\%$ error) of all the conversion factors involved in the conversion of the raw electronic measurements into the position of the RNA polymerase [47]. Various practical limitations generally make it difficult to achieve the required accuracy (section II.2.1).

In order to maximize the accuracy of our measurements, we need to impose additional structure onto the data. Specifically, following the ideas of Herbert et al. [9], we take advantage of the dependency of the velocity of RNA polymerase on the underlying sequence, by performing our measurements on a template with a repetitive sequence. On such a template, the velocity and pausing of the enzyme likewise exhibit a repetitive pattern, and the period of this pattern can be extracted with high accuracy, thus providing an end-to-end calibration method that bypasses most sources of uncertainty (section II.2.1).

The high accuracy we achieved is useful for the study of sequence-dependent dynamics. In particular, we show that the sequence dependence of backtracking (one of the mechanisms that leads to pausing) differs from the general sequence-dependence of pausing (section II.3.2). This observation indicates that other mechanisms are at play in pause entry, in contradiction with earlier reports suggesting that backtracking is the sole major pausing pathway [48]. Our localization accuracy also allowed us to show (section II.3.3) that RNA secondary structures can either enhance or diminish pausing, depending on the underlying sequence. While the effect of individual hairpins on pausing is well known [49], earlier optical tweezers studies found that the general, long-scale elongation kinetics did not depend on the presence or absence of secondary structures [50]. We hypothesize that such observations arose from an averaging of the positive and negative effects that we observed.

Temporal resolution

In the absence of pauses, the velocity of RNA polymerase is on the order of a few dozen bp/s; i.e., the enzyme spends less than 100 ms at each position along the template. Conversely, most pause events characterized so far have lifetimes on the order of seconds or more, due to limitations in the temporal resolution of the measurements [21]. Thus, the 100 ms to 1000 ms time scale of RNA polymerase dynamics remains poorly understood, even though it is accessible to some other techniques, such as rapid quench-flow experiments [51]. Studying events at such a fast time scale also presents statistical difficulties, as even a hypothetical enzyme that translocates at a speed of 20 bp/s without ever entering long-lived pauses

would sometimes take slower steps in the 100 ms to 1000 ms time scale, due to the intrinsic stochasticity of single-molecule dynamics. In other words, the pause-free and paused dwell time distributions overlap significantly; it becomes impossible to label with certainty a given event as paused or non-paused, and such assignments can only be made probabilitiscally. It is therefore necessary to develop statistical tools that allow us to analyze such overlapping distributions.

In section II.2.2, we present a nonlinear filtering method based on recent advances in the statistical [52] and algorithmic [53] literature (in particular, an efficient implementation of total variation denoising), which allowed us to take, once again, advantage of our preexisting knowledge regarding the relevant dynamics to extract information down to the 100 ms time scale. Then, in section II.2.3, we adapt a technique originally introduced for the analysis of power law tails [54] to our dataset. This technique allows us to nonparametrically compare our measurements at pause sites with measurements taken outside of the pause sites, while making only minimal modelling assumptions regarding these measurements.

Our methodology allowed us to estimate the efficiency with which a given sequence causes the enzyme to enter a state with slower forward translocation dynamics, i.e., the pausing efficiency of the sequence (section II.3.1). Contrary to earlier reports [55], we find that pausing efficiencies are generally high and independent of the applied force. We propose that the previously observed force dependence arose from the inability of earlier methods to observe short events and the attempt to compensate it by an inaccurate extrapolation of the long event distribution into the short event regime. The high temporal resolution we achieved also allowed us to observe the dynamics involved in the entry into the backtracked state, and to show that this entry occurs through a long-lived non-backtracked or at most single-base-pair-backtracked intermediate.

Resolution and accuracy of super-resolution microscopy

During the first half of my Ph. D., my research focused on another important single-molecule technique: super-resolution microscopy. Since the fundamental breakthrough, in 2006, that exploited the stochastic switching of fluorescent markers to move optical microscopy past the diffraction limit, down to a resolution of ~ 15 nm [56–58], this technique has been widely adopted, and used to study a host of biological systems [59]. We review the fundamental physical principles underlying super-resolution microscopy in chapter III.1.

Even more than for optical tweezers, super-resolution microscopy only lives up to its name through optimization of its ability to precisely locate a single fluorophore. Here again, this optimization occurs through advances in either the instrumentation or the data analysis methodology. We review the current state of the art in localization algorithms in chapter III.2.

Unlike classical diffraction microscopy, super-resolution methods (or, more accurately, methods based on stochastic switching) do not directly yield an image (i.e., a map of fluorescence intensities), but rather a list of fluorophore localizations. As such, different kinds of quantitative informations can be extracted from it. One important example is determining the oligomeric state of a protein complex. We participated in an effort to perform the *counting* of such complexes by super-resolution microscopy [60]; we review such quantification techniques in chapter III.3.

Finally, we note that, as for optical tweezers, there exists a duality between resolution and accuracy in super-resolution microscopy. In fact, at a local level, super-resolution microscopy is not the optical technique with the highest *resolution*: techniques such as single-molecule Förster resonance energy transfer (FRET) routinely report on conformation changes associated with distance changes in the sub-nanometer range [61]. Rather, it is the ability of super-resolution microscopy to *accurately* localize molecules across large field of views that makes it an invaluable technique in cell biology. In part IV, we present our work in combining the advantages of bimolecular fluorescence complementation—another technique with high resolution, but relatively limited accuracy—with those of super-resolution microscopy, through the development of a split photoactivatable fluorescent protein system.

Part I

Full molecular trajectories of RNA polymerase at single base-pair resolution

This work was published as

Full molecular trajectories of RNA polymerase at single base-pair resolution.

M. Righini*, A. Lee*, C. Cañari-Chumpitaz*, T. Lionberger*, R. Gabizon, Y. Coello, I. Tinoco Jr, and C. Bustamante, *Proc. Nat. Acad. Sci. USA* (2018), *115*, 1285–1291.

It has been reproduced here, with modifications, with the permission of all authors.[1]

This work was supported by the Howard Hughes Medical Institute (Carlos Bustamante), NIH Grants R01GM032543 and R01GM071552 (both to Carlos Bustamante), and the US Department of Energy Office of Basic Energy Sciences Nanomachine Program under Contract DE-AC02-05CH11231 (to Carlos Bustamante).

This work is dedicated to the memory of our friend and colleague Ignacio Tinoco Jr., whose example and guidance inspired our research efforts. We thank Matthew J. Comstock and Charles Wickersham for advice and help in setting up the time-shared trap, and the members of the Bustamante lab for helpful discussions.

In recent years, highly stable optical tweezers systems have enabled the characterization of the dynamics of molecular motors at very high resolution. However, the motion of many motors with angstrom-scale dynamics cannot be consistently resolved due to poor signal-to-noise ratio. Using an acousto-optic deflector to generate a "time-shared" dual-optical trap, we decreased low-frequency noise by more than one order of magnitude compared with conventional dual-trap optical tweezers. Using this instrument, we implemented a protocol that synthesizes single base-pair trajectories, which are used to test a Large state-space hidden Markov model algorithm to recover their individual steps. We then used this algorithm on real transcription data obtained in the same instrument to fully uncover the molecular trajectories of *E. coli* RNA polymerase. We applied this procedure to reveal the effect of pyrophosphate on the distribution of dwell times between consecutive polymerase steps.

[1] PNAS permits an author to reuse articles for their own dissertation.

Chapter I.1

Introduction

Proteins involved in a wide array of cellular functions are able to convert chemical energy into mechanical motion, thus functioning as molecular motors [30]. A comprehensive description of the dynamics of such motors requires following their position with sufficient spatiotemporal resolution, i.e., to determine their molecular trajectory. The trajectories of all motors described to date consist of alternating stationary periods (known as "dwells") and translocation events (known as "bursts"). From these trajectories, we can extract fundamental parameters of a motor's dynamic operation, such as the distribution of its step sizes and dwell times; these parameters, in turn, provide crucial insight into the mechanochemical coupling underlying the motor's operation. For motors, such as dynein, take steps with variable sizes [36], characterization of the molecular trajectory reveals how the motor adapts its step size to the conditions under which it operates (external load, ATP concentration, crowded environment, etc.). Conversely, knowledge of the dwell time distribution can, for example, shed light on the coordination mechanism in multi-subunit motors [43, 62–66].

Optical trapping can be used to characterize molecular motor dynamics with high precision over biologically relevant times, distances, and forces. The molecular trajectories of motors that take relatively large steps (such as kinesin, which takes 8 nm steps on microtubules) are now regularly accessed in many laboratories. However, the ability to reliably and routinely resolve the molecular trajectories (including all steps and inter-step dwell times) of many nucleic acid-associated motors (e.g., DNA and RNA polymerases, helicases, dsDNA translocases, etc.), whose steps are on the order of 1 bp ($\sim$ 3.4 Å), continues to elude biophysicists. While base-pair stepping by RNA polymerase and helicases have been previously observed with optical tweezers occasionally and over short distances and time scales [39, 67], sufficiently low levels of instrumentation noise even in the most sophisticated instruments are short-lived (typically lasting on the order of tens of seconds) and infrequent enough that upwards of 90 % of the data have to be ignored and discarded [39]. Thus, extraction of molecular trajectories with single base-pair resolution in a reliable and consistent way has not been possible until now.

Here, we compare the resolution of two optical trap designs: split-path and time-shared optical tweezers instruments under identical conditions. We show that the ability to robustly extract trajectories with single base-pair resolution is limited by low-frequency noise present in the split-path design, but not in the time-shared design. We introduce a protocol to experimentally synthesize trajectories simulating single base-pair stepping by a molecular motor. The synthesized data are used to evaluate the fitness of the tether and to test the performance of a large state-space hidden Markov model (LSS-HMM) algorithm in extracting the corresponding molecular trajectories. Finally, we use this same algorithm to extract the full molecular trajectories (steps and dwells) of *E. coli* RNA polymerase from transcription traces obtained in the time-shared instrument and to characterize the effect of pyrophosphate (PPi) on the distribution of dwell times between steps of the enzyme.

Chapter I.2

Theory

The fluctuations of a microscopic bead held in a harmonic trap of stiffness k are described by its power spectrum. According to the fluctuation-dissipation theorem, the random, uncorrelated forces due to the collisions of surrounding molecules (at a temperature T) give rise to a Lorentzian power spectrum for the position of the bead (in the strongly overdamped regime) [68],

$$S_x(f) = \frac{k_B T}{\pi^2 \gamma (f^2 + f_c^2)} \,. \tag{I.1}$$

Here, γ is the drag coefficient, k_B the Boltzmann constant, and $f_c = k/2\pi\gamma$ the corner frequency, beyond which the system begins to lag behind an external driving stimulus (typically in the kilohertz range for optical traps). Equation I.1 describes how the noise is distributed over frequencies: the spectrum of fluctuations is approximately flat (white noise) at frequencies $f < f_c$ and decreases as $1/f^2$ for $f > f_c$.

For a measurement at a bandwidth B, the mean quadratic displacement of the trapped bead, $\left\langle \Delta x^2 \right\rangle$ can be computed by integrating the power spectrum $S_x(f)$ of the trajectory over frequencies ranging from zero to B,

$$\left\langle \Delta x^2 \right\rangle_B = \int_0^B S(f)\, \mathrm{d}f, \tag{I.2}$$

which yields in the limit of low bandwidth

$$\left\langle \Delta x^2 \right\rangle_{B \ll f_c} = \frac{2 k_B T}{\pi k} \frac{B}{f_c} \tag{I.3}$$

and in the limit of high bandwidth

$$\left\langle \Delta x^2 \right\rangle_{B \gg f_c} = \frac{k_B T}{k} \,. \tag{I.4}$$

The latter result is known as the equipartition theorem.

Ultimately, the quantity that determines whether a change in displacement of the bead Δx (due, for example, to the displacement of a molecular motor) can be distinguished from all other fluctuation sources is the signal to noise ratio (S/N)—that is, the ratio of this extension change to the root-mean-square displacement of the bead,

$$\frac{S}{N} = \frac{\Delta x^2}{\sqrt{\langle \Delta x^2 \rangle}}\,. \tag{I.5}$$

Replacing the mean-square displacements obtained in the limits of low and high bandwidths, the signal-to-noise integrated to bandwidth B is

$$\left(\frac{S}{N}\right)_{B \ll f_c} = \frac{\Delta F}{\sqrt{2\gamma B k_B T}} \quad \text{and} \quad \left(\frac{S}{N}\right)_{B \gg f_c} = \frac{\Delta F}{\sqrt{2 k_B T k}} \tag{I.6}$$

where $\Delta F = k\Delta x$ is the change in tether tension due to the bead displacement.

Thus, in principle, even very small displacements can be observed with a signal-to-noise ratio greater than one simply by decreasing the bandwidth B to well below the corner frequency, provided that the instrument is Brownian noise-limited and that the bandwidth does not compromise the temporal resolution of the experiment. As discussed below, the first of these conditions is rarely fulfilled.

Chapter I.3

Differential path dual-trap configuration

A common assay to record the position of a motor as a function of time is to optically trap a bead that is linked to the motor's substrate (e.g., a microtubule, a DNA template, etc.), while the motor itself is directly or indirectly attached to another bead held in a second trap [38, 69, 70]. In this dual-trap assay, the progress of the motor along its track is reported by the distance between the beads (differential detection scheme) [38]. Any correlated motion of the two traps does not change the distance between the beads and is thus automatically removed when calculating the trap-to-trap distance. Only anticorrelated motion can contribute to the measured signal [38].

The most common dual-trap configuration is known as the split-path geometry [38, 71]. There, the two traps are generated by splitting using a polarizing beam splitter the polarized light of a laser source (in our case, an Nd:YAG 1064 nm) into two beams that travel through different paths, one of which includes a piezo-actuated mirror for steering the beam, until they are recombined in a second polarizing beam splitter, slightly shifted in angle relative to one another and finally sent through the back focal plane of a focusing objective (figure I.1a) and focused into home-made fluidics chambers. Detection of bead positions and forces was achieved by collecting the light on a second objective, splitting the beams again using a polarizing beam splitter, and imaging the beams on quadrant photodiodes (QPDs).

The noise in a split-path dual-trap instrument can be determined by tethering a DNA molecule between two beads held in the traps (figure I.2A) and by monitoring their net differential displacement over time under applied tensions. The total noise to infinite bandwidth is the sum of correlated and anticorrelated contributions; moreover, its value only depends on the temperature and the combined stiffness of the tether and the traps (equation I.4). As the tension is increased, the tether stiffens (due to the nonlinear mechanical properties of the DNA) and the beads' motions become increasingly correlated. Consequently, the anticorrelated component (which is the only one measured in differential detection) must necessarily decrease, both because the total noise decreases and because a larger fraction of it goes in the correlated component (figure I.2B) [72].

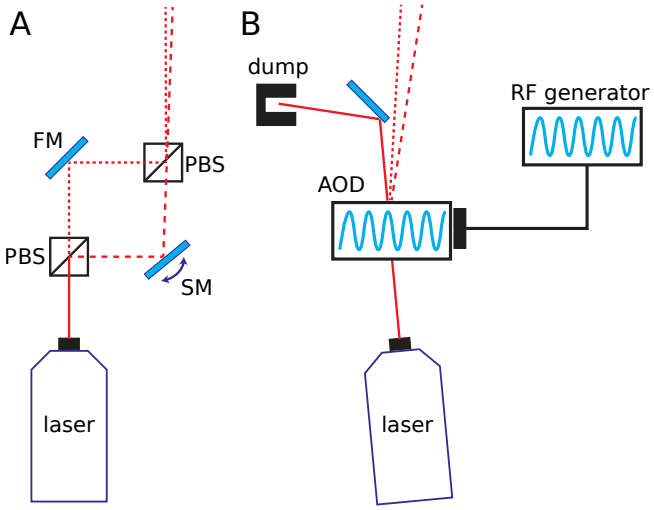


Figure I.1:
Simplified diagrams of the optical setups.
(A) The split-path setup splits the laser light into two orthogonally polarizing beams to steer independently one trap (FM, fixed mirror; PBS, polarizing beam splitter; SM, steerable mirror).
(B) Time-sharing the traps with an AOD eliminates the need for the split paths.

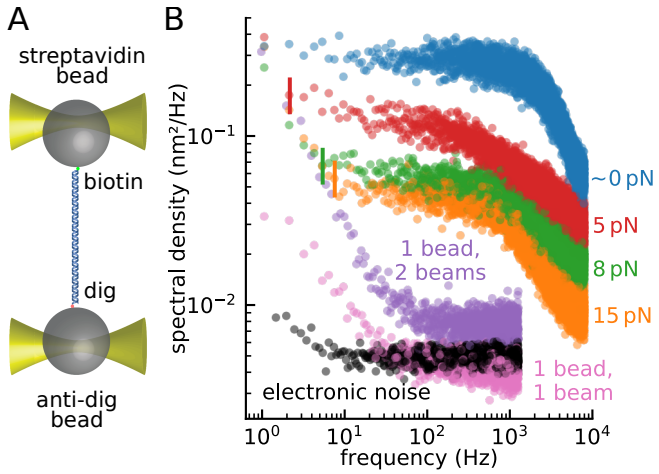


Figure I.2:

Measurement of the noise in a split-path instrument.

(A) Two 1 µm beads were tethered by 1 kb DNA, using biotin-streptavidin linkage on one bead and digoxigenin-antidigoxigenin on the other. The beads were trapped using a split-path dual trap.

(B) Power spectra of the differential signal were recorded while the tether was held under various tensions: ~ 0 pN (blue), 5 pN (red), 8 pN (green), and 15 pN (orange). The vertical lines indicate the frequencies below which a non-Lorentzian component emerges from the Brownian floor. The purple curve shows the power spectrum of the differential signal from a single bead trapped with both trapping beams. The single-bead measurement measures relative drift between the two optical traps as well as contributions from bead-related artifacts (pink) and electronic noise (black).

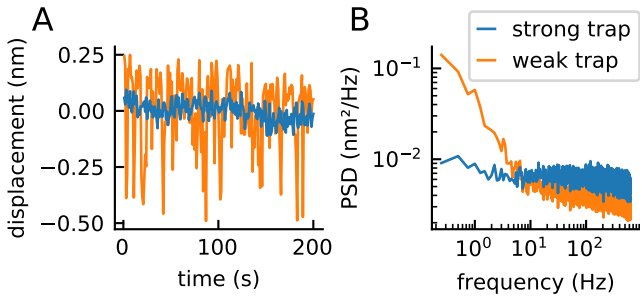


Figure I.3:

Measurement of the noise of a bead trapped by two colocalized beams in a split-path instrument.

(A) A strong and a weak trap were formed in a split-path instrument and were focused onto the same bead. The recorded displacement signal in the two traps shows that the vast majority of the noise occurs in the weak trap signal, suggesting that the anticorrelated signal arises because the two traps are physically drifting relative to each other.

(B) The power spectrum of the weak trap signal measured in (A) exhibits low-frequency pink noise comparable to the one measured on the tethered construct (figure I.2), whereas no pink noise component is observable in the power spectrum of the strong trap signal. Similarly as in (A), this suggests that the increased low-frequency noise in the differential signal is due to relative trap drift.

Note, however, that as the anticorrelated noise is suppressed, a non-Lorentzian noise source becomes apparent at lower frequencies. Because this noise is independent of force, the frequency at which it emerges over the Brownian floor becomes larger as force increases (see vertical lines in figure I.2B). Although the power spectrum of fluctuations of a bead in a single trap is almost white below the corner frequency (figure I.3B, blue curve) [73], the power spectra of two tethered beads in all dual traps display this low-frequency $1/f$ noise component [39, 71, 73–75], also known as pink noise. Equation I.6 indicates that resolving single base-pair stepping by RNA polymerase requires using a bandwidth as low as a few hertz (although the exact value depends on the signal-to-noise ratio required for detection, which depends itself on the algorithm used); at this bandwidth, the low-frequency noise, rather than Brownian noise, becomes the resolution-limiting factor.

Several sources for this low-frequency noise have been proposed, including optical turbulence in the split paths, trap positional instability, bead asymmetry, and tether dynamics [39, 71]. Indeed, this noise can be reduced—but not eliminated—by replacing air with helium in the optical path or by shortening the length of the split paths [39, 71]. However, the origin of the residual noise remains unknown. Here, we establish that a major contributor to

the low-frequency noise is the time-dependent change (physical drift) of the trap positions relative to one another.

To characterize the positional stability of the traps in our split-path instrument, we focused both traps onto a single 1 μm-diameter microsphere and monitored the differential signal (Δx). In this dual-beam, single-bead experiment, the differential signal only reports the relative trap displacements at the focal plane (figure I.2B, purple curve). Note that the low-frequency noise remains present even without a tethered molecule or without any protein coating on the microspheres. Thus, neither tether attachment dynamics nor excess molecules bound to the beads can fully account for the $1/f$ noise. Notice that the electronic noise floor (figure I.2B, black curve) is well below the single-bead noise.

In these single-bead experiments, the single trap formed by precisely overlaying the two orthogonally polarized trapping beams emerging from the split path is equivalent to a single beam polarized at 45°. Strikingly, however, when a single 45° polarized beam is focused to trap the bead, the low-frequency noise is reduced significantly (figure I.2B, pink curve) compared with the same measurement using the beam-steering path (figure I.2B, purple curve). Finally, focusing a weak and a strong trap onto the same bead and measuring the fluctuations in both channels shows that nearly all of the measured drift occurs in the weaker trap, demonstrating that the anticorrelated signal arises because the two traps physically drift relative to each other (figure I.3) (drift is mainly encoded into the weak trap since the bead tends to follow the stronger trap). Thus, we conclude that the positional instability of the two traps originates within the beam-steering path of the optical trap and is the main source of the low-frequency noise.

The remaining low-frequency noise above the electronic floor (figure I.2B, compare pink and black curves) could originate from asymmetry or optical anisotropy of the beads. However, this errant displacement signal (pink curve) contributes an insignificant amount of noise compared with the single bead measurement when the instrument includes the beam-steering path (purple curve).

What causes the positional instability of the traps in the split-path design? Since the positions of the traps depend directly on the angles at which the beams enter the back focal plane of the objective lens, the relative positional stability of the traps is determined by the relative angular stability of the respective beams. To move one trap relative to the other in the split-path design, the orthogonally polarized beams are steered immediately after they are separated and before they are recombined (figure I.1A) [39, 71, 73–75]. The optical components in each path exclusively interact with one of the beams and thus can introduce angular drift between them. Therefore, an alternative way of steering the traps that does not require the light path to be branched can overcome this limitation.

Chapter I.4

Time-shared dual-trap configuration

A single beam can be used to form two traps, if its direction is switched at a high rate by an acousto-optic deflector (AOD) [40, 70, 76–80]. Two beads, linked by a DNA tether, can be trapped in this system, whose corner frequency is, as before, $f_c = k/2\pi\gamma$, where k is the total stiffness of the DNA and one of the traps (as only one trap is on at any time). This corner frequency (typically in the kilohertz range) determines the rate above which the AOD must switch the beam direction to keep the beads stationary. In our case, this switching occurred every 5 μs and was controlled by a custom-made radio-frequency board.

This time-sharing scheme eliminates the need to split the light into two different paths (figure I.1B) while maintaining the ability to individually steer each trap by controlling the amplitude of the deflection through the AOD. In this scheme, the beams forming each trap never encounter different optical components, and any mechanical drift of these components can only result in a correlated motion of the two traps that is automatically cancelled when calculating the distance Δx between the beads in the traps.

We compared the steady-state stability of a split-path dual-trap and a time-shared dual-trap (both custom-built) under identical conditions, as follows. We linked two polystyrene beads of 1 μm-diameter with a 3.5 kb dsDNA by means of streptavidin/biotin and digoxigenin/antidigoxigenin conjugations and held the tether for 5 min under 5 pN of tension. The power spectrum of the differential signal in the time-shared instrument reveals a low-frequency noise more than one order of magnitude smaller than that of the split-path instrument at 0.1 Hz (figure I.4). A similar conclusion was arrived at using an alternative split-path instrument. This result supports the idea that the relative drift of the split beams is responsible for the low-frequency noise.

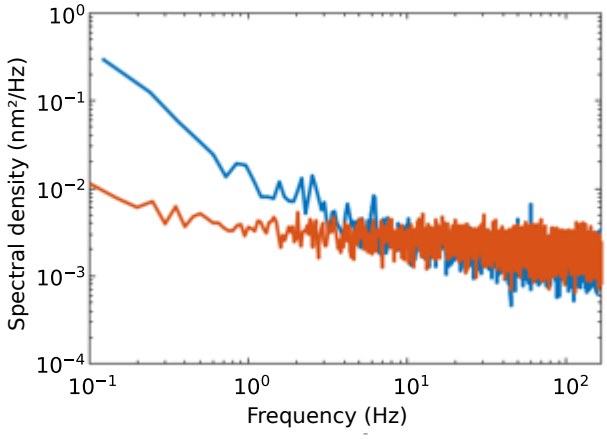


Figure I.4:
Power spectra of a split-path (blue) and a time-shared (red) instrument obtained monitoring the extension of a 3.5-kb DNA tether held at 5 pN of tension. $1/f$ noise dominates at low frequencies, especially in the split-path instrument. Only the portion of the spectra below the corner frequency ($f < f_c$) is shown.

Chapter I.5

Large state-space hidden Markov model algorithm

Most motors take steps of a single, constant size. For such motors, stepping motion may be resolved even if the signal-to-noise ratio is locally smaller than unity. For example, if a motor sometimes takes steps that are sufficiently slow, the step size distribution can be estimated from such slow regions and then be used to elucidate the stepping motion over regions of faster displacement.

Specifically, we adapted the LSS-HMM fitter [45, 81, 82], which models the measured trace as arising from a random process, as follows. At any time, the molecular motor is assumed to occupy an unknown position (the "hidden state"), discretized to a small "state size" chosen to be much smaller than the expected step size of the motor—we set it to 0.025 nm (less than 0.1 bp). The position measured by the optical tweezers is modeled as the sum of the actual position of the motor and of a Gaussian error with an unknown but fixed variance s^2. Between each time point, the motor moves by a random amount (zero if the motor is not moving); the size d of this displacement (the "step size") is drawn from an unknown but fixed distribution, $p(d)$.

The procedure to find both the step-size distribution $p(d)$ and the noise variance s^2 that maximizes the likelihood of observing the trace that was actually measured is called the Baum-Welch algorithm. However, due to the large number of states in the model, a specific optimization (introduced by Felzenzswalb et al.), must be used [45]. Unlike most other popular approaches to step-finding [83, 84], the LSS-HMM algorithm learns the distribution of step sizes from the data and can therefore avoid taking large jumps upon encountering an outlier in the trace, all without manual intervention.

We implemented the LSS-HMM algorithm in the Cython programming language [85]. The analysis parameters (described in the above-mentioned references) were set as follows: experimentally synthesized traces (chapter I.6) and segments of transcription activity traces between consecutive adjustments to the trap position (chapter I.7) were downsampled to

200 Hz; the quantization size was set to 0.025 nm; a null prior was used for negative step sizes (for reasons explained in chapter I.7); each HMM was run for up to 1000 iterations, up to 60 seconds, or until the change in likelihood between iterations dropped below 10^{-8} (whichever occurred first).

Chapter I.6

Experimental synthesis of stepping trajectories

To characterize the resolution capabilities of the time-shared dual-trap, we implemented a procedure using a tethered molecule to experimentally synthesize the molecular trajectories of a motor taking single base-pair steps according to a prespecified dwell time distribution (figure I.5). We call this procedure "STEPS", for Stepping Trajectories by ExPerimental Synthesis.

To fully reproduce the noise characteristics present in transcription elongation, the STEPS procedure was performed using a stalled elongation complex tethered via a DNA handle to a 1 µm-diameter bead (held in one trap), while the distal end of the DNA template was tethered to another 1 µm bead (in the second trap), yielding a 3.5 kb tether kept under 15 pN of tension.

Briefly, stalled complexes were prepared by incubating 2 nM DNA with 10 nm RNAP in TB20 (Tris 20 mM pH = 8, 20 mM NaCl, 20 mM DTT, 10 mM $MgCl_2$, 20 µg/ml casein) for 20 minutes at 37 °C. The complexes were then ligated to the beads at a ratio of 1 fmol stalled complex to 2 µg beads in TB20 in the presence of 0.1 mM ATP and 0.4 unit of T4 DNA ligase, for 60 minutes at room temperature. For 1.5 kb DNA handles, 1 fmol handle was ligated to 3 µg beads. Following the ligation, heparin was added to 0.4 mg/ml to the beads. To the beads ligated to the DNA handle, a 200-fold excess of neutravidin was added and incubated with the beads for 10 minutes prior to diluting with experimental buffer, namely, Tris 20 mM pH = 8, 130 mM KCl, 10 mM $MgCl_2$, 0.1 mM DTT, 0.1 mM EDTA and 10 mM NaN_3 (added as a singlet oxygen scavenger to reduce the extent of photodamage [86]). For the stalled complex beads, beads were incubated for 10 minutes with the added heparin before dilution with experimental buffer.

In this geometry, polymerase translocation would cause a corresponding change in bead-to-bead distance. After obtaining a power spectrum of the beads, for calibration of the conversion factors between measured voltages and bead displacements, and after forming a

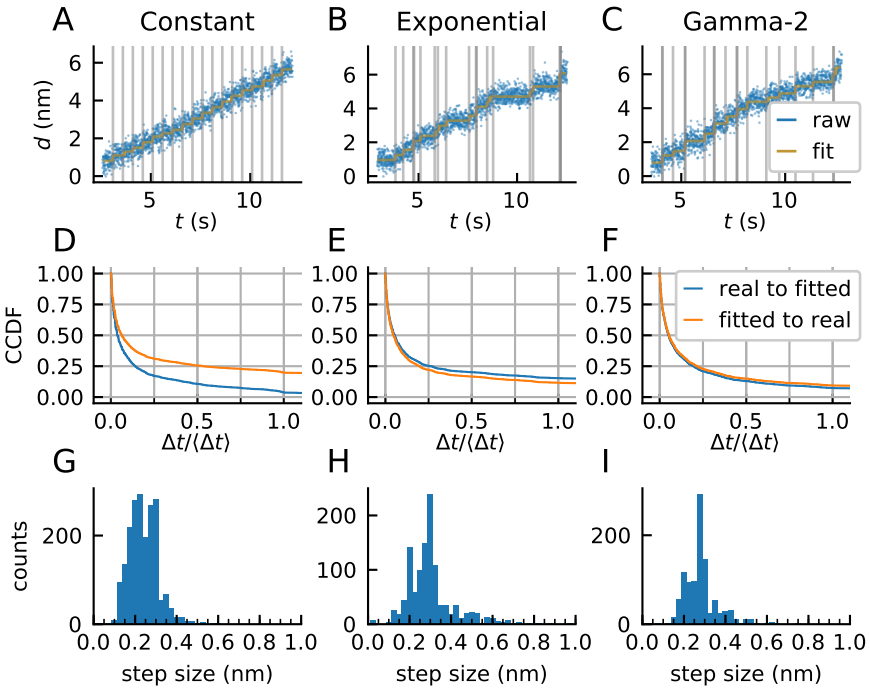


Figure I.5:
Collection and analysis of STEPS data. (A–C) Traps holding tethered beads were displaced away from each other by 0.34 nm at times separated by (A) constant, (B) exponentially distributed, or (C) gamma-2 distributed dwells, in each case with a mean dwell time of $\Delta t = 0.5\,\mathrm{s}$ (blue, differential signal at 200 Hz; gray vertical lines, times of actual trap motion), and fitted by LSS-HMM without knowledge of the step size or the stepping times (gold, result of the fit). Good agreement between fitted times and actual trap motion can be observed. (D–F) Blue, complementary cumulative distributions of the time intervals between each real step and the closest fitted step, with the constraint that two real steps may not be associated with the same fitted step. Orange, complementary cumulative distributions of the time intervals between each fitted step and the closest real step, with the constraint that two fitted steps may not be associated with the same real step. The time axis is in units of mean dwell time ($\langle \Delta t \rangle = 0.5\,\mathrm{s}$). (D) Constant, (E) exponentially distributed, and (F) gamma-2 distributed dwells. (G–I) Step size distribution in the three cases. (G) Constant, (H) exponentially distributed, and (I) gamma-2 distributed dwells.

tether (by rubbing the beads against each other), we thus simulated such a motion by moving one trap toward the other in twenty 0.34 nm (1 bp) increments according to various time interval protocols (dwells): constant dwells (figure I.5A), exponentially distributed dwells (figure I.5B), and dwells drawn from a gamma distribution with shape parameter 2 ("gamma-2", figure I.5C). The second and third cases simulate a molecular motor translocating with one and two rate-limiting steps, respectively.

In classical optical tweezers measurements, the distances of the beads to the centers of their respective traps, $\Delta x_{\text{bead1-trap1}}$ and $\Delta x_{\text{bead2-trap2}}$, are subtracted from the distance between the traps, $\Delta x_{\text{trap1-trap2}}$ (the latter being set by the experimentalist). Such an approach is suitable to track the time-varying extension of the tether between the traps. In the case of STEPS data, however, because the tether is $\sim$ 17× stiffer than the traps (1.7 pN/nm at 15 pN vs two traps of 0.2 pN/nm each), the tether extension only changes by 1/18 of the trap motion, whereas the total displacement of the beads away from the traps changes by 17/18 of this motion (i.e., 0.32 nm as the traps were moved by 0.34 nm). Thus, we chose to analyze the total displacement data.

Recorded STEPS traces were fitted using the LSS-HMM algorithm, without prior knowledge of the actual step size and trap motion times. Traces where the LSS-HMM algorithm failed to converge (usually, due to a numerical underflow of the likelihood) were discarded.

To quantify the time accuracy of the fit for STEPS data, we first determined if all real steps (i.e., actual trap motions) were correctly detected, or if some of them were missed (underfitting). Likewise, we evaluated if scored steps, as scored by the LSS-HMM algorithm, correspond to real steps, or some of them were spurious (overfitting).

In order to quantify underfitting, we paired each real step with the closest scored step. If done naively, such pairing could fail to detect a case where two temporally close real steps are fitted with a single step—thus missing a short dwell between the two real steps. Therefore, we additionally imposed the condition that the pairing between fitted steps and real steps must be one-to-one (if there the number of fitted steps was smaller than real steps, some real steps were left unpaired). Specifically, we required that the number of matches be equal to the lower between the number of real steps and scored steps, and minimized the sum of the time intervals between the paired steps.

We then ask, what is the distribution of the time intervals between the real and the scored step in the pairings? In other words, how far is each real step from the closest scored step? We find that regardless of the dwell time distribution (constant, exponential or gamma-2), more than 70 % of the real steps are within 100 ms (one fifth of the mean dwell time) of the closest scored step (figure I.5d-f).

In order to quantify overfitting, we asked how far each scored step is from the closest real step, once again imposing one-to-one correspondence. In this case too, at least 70 % of the scored steps were found within 100 ms of the closest real step, with the exception of the

biologically less relevant case of constant steps, where, by experimental design, the allowance had to be raised to 165 ms (figure I.5g-i).

On the other hand, the step size distributions are relatively broad in all cases (constant stepping: (0.24 ± 0.25) nm; exponential: (0.30 ± 0.25) nm; and gamma-2: (0.27 ± 0.25) nm; all values are mean $\pm$ standard deviation) (figure I.5G–I).

Thus, interestingly, step times can be correctly obtained even though the step sizes are recovered with limited accuracy. We rationalized this observation on the basis that LSS-HMM can assign variation in the measured bead position to two sources—actual spread of the step size distribution and additional Gaussian noise; thus, LSS-HMM can choose to report a wider step size distribution in order to narrow the Gaussian noise distribution. We note, however, that the width of the step size distribution is not broad in absolute terms compared with step size distributions seen for other motors—it only appears so here because the step size is small compared with the magnitude of the noise.

Chapter I.7

Full experimental trajectories of RNA polymerase

Analysis of STEPS data allowed us to establish that the time-shared dual-trap instrument used in combination with the LSS-HMM algorithm can score single base-pair steps from synthesized data in an accurate and robust manner. We could now take up the challenge of extracting the full molecular trajectories of individual elongating *E. coli* RNA polymerase molecules, with base pair resolution and over long distances.

Tethered, stalled elongation complexes were prepared and trapped as described above. Once a stalled elongation complex passed the test of the STEPS protocol (a total of six times—twice for each dwell time distribution constant, exponential, and gamma)), we then delivered 10 μM NTP (using a home-made flow-chamber with separate entries for buffer and for the NTP solution) and recorded the tether extension under a mean tension of 15 pN applied in a direction that assisted forward translocation (figure I.6A). In this geometry, translocation by the polymerase causes the tension to decrease; whenever the tension dropped below 14 pN, the traps were displaced to restore a tension of 16 pN ("semi-passive" mode). At the low NTP concentration of 10 μM, the pause-free transcription velocity is three times slower than the rate used in the STEPS procedure (0.62 bp/s vs. 2 bp/s), and single steps can be resolved. We discarded any trace where the total change in tether extension was less than 5 nm or where extraneous noise was visually obvious. All other traces, amounting to 14 625 s of activity covering a distance of 1589 nm, were taken into account for further analysis.

The theory of the LSS-HMM algorithm [45, 81, 82] allows the molecular motor to take steps both in the forward and in the backward direction (backtracking). However, under 15 pN of assisting force, backtracking events are rare; we found it beneficial to force the steps to be always in the forward direction. Specifically, the prior on the step size distribution $p(d)$ was set to zero for $d < 0$. In the absence of such a constraint, some of the traces would be fitted to a collection of steps that quickly alternate between forward and backward motion (about every 100 ms). The origin of this motion is not known but it may be due to dynamics

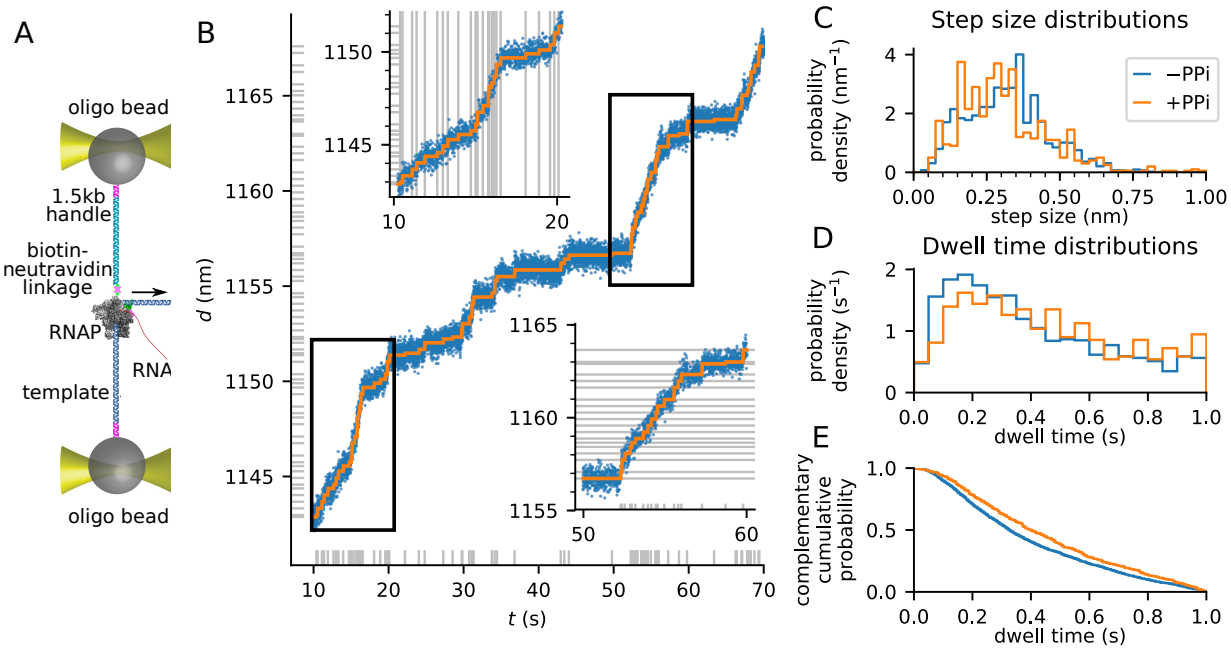


Figure I.6:

Observation of full molecular trajectories of RNA polymerase at single base-pair resolution.

(A) Experimental geometry used to record transcription activity under assisting force, on a time-shared instrument, as described in the main text.

(B) A sample transcription trace, covering 70 bp (blue, raw data) in 1 min, and the fitted molecular trajectory (orange). Horizontal gray ticks indicate the positions of the dwells between two steps. Vertical gray ticks indicate the times of the scored steps. Insets are zooms into the regions marked by black rectangles. Horizontal gray ticks or lines in insets mark the position of each fitted step, showing that they are separated by ~ 0.33 nm. Vertical gray ticks or lines indicate the times of the scored steps.

(C) The distribution of fitted step sizes in the absence of PPi (−PPi, blue) is peaked at (0.32 ± 0.15) nm, corresponding to the expected size of 1 bp. The distribution of fitted step sizes in the presence of 100 µM PPi (+PPi, orange) is peaked at (0.30 ± 0.15) nm, corresponding to the expected size of 1 bp.

(D) The distributions of dwell times in the absence and presence of PPi are not exponential, but peaked around $\Delta t \sim 0.2$ s.

(E) Comparison of the complementary cumulative dwell time distributions in the absence (blue) and presence (orange) of PPi shows that PPi slows down processive elongation by 20 %.

occurring at the tethering point of the RNAP. Preventing any backwards motion forces the LSS-HMM to average out these dynamics into a single state.

In certain cases, the LSS-HMM would fail to converge to a valid step distribution. As for STEPS data, such a failure usually manifests itself as a numerical underflow of the likelihood. Such failures could arise due to two reasons. First, a trace can exhibit actual backtracking, contradicting our initial assumption. Second, RNA polymerase can enter long-lived pauses [87, 88] that we also found to be, counter-intuitively, detrimental to the performance of the LSS-HMM—likely due to the difficulty for the LSS-HMM algorithm to distinguish between very slow activity and residual low-frequency noise.

In such cases, we split the data into two halves that were fitted separately, and the procedure was repeated until the fit succeeded, the segment length dropped below 5 s, or the total transcribed distance dropped below 10 s (the analysis of segments containing very few steps is difficult for any HMM-based algorithm, as it needs to *learn* the correct step distribution from the dataset self-consistently). In such manner, periods of backtracking or of long-lived pauses can be separated from the analysis, whereas segments containing processive activity would be analysed. Note that if the abnormally noisy traces described above had not been removed manually, this algorithm would likewise reject them.

Overall, the data successfully fitted by LSS-HMM amounted to $N = 3874$ steps from 30 different molecules, corresponding to 6344 s of activity covering 1198 nm. Despite the lack of prior assumption on the step size, the histogram of observed step sizes peaks at (0.32 ± 0.15) nm (figure I.6C), which compares favorably with the expected step size of 0.33 nm at 15 pN as predicted by the worm-like chain equation [72]. We were able to recover the single base-pair trajectory with high accuracy in segments as long as 70 bp (figure I.6B). Note that the maximum length of the fitted segment is due to the need to maintain the force within a range of 2 pN; with a trap stiffness of 0.2 pN/nm per trap, the traps must be displaced—and thus a new fit region must be started—every 20 nm (10 nm on each trap), i.e., approximately every 70 bp.

The fitted segments amounted to 75 % of the total distance transcribed, but only 43 % of the total duration of the traces, due to the selective removal of long-lived pauses from the analyzed datasets. As such, the dwell time distribution obtained from the analysis faithfully represents the true distribution of pause-free translocation in short time scales; however, the distribution is underestimated at longer time scales due to the rejection of slow segments. Note that the segments rejected are well-defined, and could be subjected to further analysis by pooling short segments together before fitting with LSS-HMM in order to increase the statistics. Our results should be contrasted with earlier reports of the observation of single base-pair stepping in optical tweezers, which was limited to short segments (~ 15 bp) corresponding to ~ 10 % of the distance transcribed in ~ 10 % of the collected traces [39].

Finally, we tested the effect of PPi on the dynamics of RNA polymerase at the base-pair scale. In the presence of PPi, RNA polymerase can catalyze the pyrophosphorolysis of the

nascent RNA chain [89]. Under intermediate concentrations of PPi, transcription elongation thus occurs at a reduced rate [90]. We collected transcription traces in the presence of 100 µM PPi. From 6730 s of activity, covering a distance of 470 nm, we successfully fitted, using the same procedure as above, $N = 800$ steps (51 % of the total distance) over 1509 s of activity (24 % of the total duration), with a step size of (0.30 ± 0.15) nm (figure I.6C).

Visual inspection of the traces did not reveal significant long-lived pausing events due to the binding of PPi. On the other hand, comparison of the dwell time distributions of the transcribing enzyme in the absence and presence of PPi (figure I.6 D and E), truncated to events shorter than 1 s due to the underestimation of long events noted above, showed that the enzyme slows down due to a 20 % lengthening of the median dwell periods during processive elongation, from 0.33 s to 0.40 s (Mann-Whitney test, $p < 10^{-4}$). The dwell time distributions were not exponential in either case but peaked around $\Delta t \approx 0.2$ s.

Chapter I.8

Conclusion

Optical tweezers is a powerful method to investigate the dynamics of molecular motors. These dynamics are encoded in the interspersed dwells and steps of the molecular trajectories of a motor. The partitioning between these phases and its dependence on various external conditions provide important information about the motor mechanism. However, extracting the full molecular trajectories of motors such as RNA and DNA polymerases, helicases, and other translocases with step sizes of one base pair has been challenging. The spatial resolution of optical tweezers can be improved by increasing the tension applied on the tether (if the motor can remain active under such tension), by shortening or stiffening the handles [91], or by using smaller beads [38]. However, most optical tweezer instruments display a $1/f$ noise component that greatly limits their resolution in the frequency range where motor dynamics are monitored. We identified a source of low-frequency noise in the split paths of the most common dual-trap configuration. By using a time-shared scheme, we eliminated path splitting and decreased low-frequency noise more than ten-fold. We also implemented a protocol (STEPS) that allows us to directly check the quality of a tether in real time, before the addition of nucleotide triphosphates, and test the performance of the LSS-HMM step-finding algorithm. Altogether, these improvements permitted us to fully uncover the molecular trajectories of *E. coli* RNA polymerase at single base-pair resolution in a robust and consistent manner. We have demonstrated the power of this approach by measuring the effect of PPi on the dwell time distribution of actively elongating polymerases.

The ability to resolve single base-pair stepping and the interspersed dwells in a reliable manner and over large distances in optical tweezers opens the possibility to study the sub-nanometer activity of many molecular motors. For instance, the effect of mutations or antibiotics on the molecular trajectories of RNA polymerase can now be resolved in terms of the phases of the motor's cycle. Similarly, the ability to precisely follow the enzyme dynamics upon each nucleotide incorporation will make it possible to investigate how the template sequence controls transcription elongation and characterize the dynamics of other important

processes such as transcription proofreading [87, 88, 92], termination [93–95], or transcription through the nucleosome [96–98] with unprecedented detail.

Chapter I.9

Materials and methods

All DNA modifying enzymes were purchased from New England Biolabs. Oligonucleotides were purchased from IDT. Nucleotide triphosphates were purchased from Thermo Scientific, and standard salts and buffer components were purchased from Sigma Aldrich.

I.9.1 Polystyrene beads

For power spectra measurements, we used 0.81 µm-diameter streptavidin-coated polystyrene beads (Spherotech), and 1 µm-diameter carboxylated beads coated with anti-digoxigenin antibody (Roche, catalog number 11333089001), as follows:

10 µl of 10 % bead suspension was washed with coupling buffer (MES 0.1 M pH = 4.7, 150 mM NaCl) 4 times, with centrifugation steps (5 minutes at 4500 × g) between the washes. The beads were dispersed in 500 µl coupling buffer. 10 µl of 3.5 mM 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) were added, followed by 65 µl of 0.2 mg/ml of antibody solution. Reaction proceeded overnight at 4 °C. At this point 10 µl of 1 M glycine was added, and Tween 20 was added to 0.05 %. The beads were centrifuged, and washed 5 times with storage buffer (Tris 20 mM pH = 8, 130 mM KCl, 0.05 % Tween 20, 1 mM EDTA, 5 mM sodium azide) with 3 minute sonication steps between washes. Beads were stored at 4 °C at a concentration of 1 % w/v until use.

For preparing oligonucleotide-coated beads, the following oligonucleotides were ordered HPLC purified and used as received:

```
Bead amine: 5' /5AmMC6/TTAATTCATTGCGTTCTGTACACG 3'
Bead CGGT:  5' /5Phos/CGGTCGTGTACAGAACGCAATGAATT 3'
Bead ACCG:  5' /5Phos/ACCGCGTGTACAGAACGCAATGAATT 3'
```

To prepare a double-stranded oligo for coupling, Bead Amine oligo was hybridized to Bead CGGT oligo or to Bead ACCG oligo to generate a double stranded oligo containing a

phosphorylated 5′ overhang. Annealing was performed by heating a 1:1 mixture of the oligos in water (0.25 mM each) to 95 °C for 10 minutes, followed by cooling to room temperature on the bench. This resulted in the following oligos:

```
CGGT duplex: 5' NH2-TTAATTCATTGCGTTCTGTACACG           3'
             3'       TTAAGTAACGCAAGACATGTGCTGGC/phos  5'

ACCG duplex: 5' NH2-TTAATTCATTGCGTTCTGTACACG           3'
             3'       TTAAGTAACGCAAGACATGTGCGCCA/phos  5'
```

1 µm-diameter carboxylated polystyrene beads (Bangs Labs) were coupled to the prepared double-stranded oligos as follows: 10 µl of 10 % (w/v) beads were washed 4 times with 200 µl coupling buffer (MES 0.1 M pH = 4.7, 150 mM NaCl, 5 % DMSO), and dispersed in 20 µl coupling buffer. All centrifugations took place for 5 minutes at 4500 × g. 10 µl of 20 µM double stranded oligo and 6 µl of 2 M 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) were added, followed by vigorous shaking for 2 hours at room temperature. At this point another 10 µl of 2 M EDC were added, followed by overnight shaking at room temperature.

The remaining active EDC was then quenched by adding 2.5 µl of 1 M glycine, and the beads were washed 5 times with storage buffer (Tris 20 mM pH = 8, 1 mM EDTA, 0.05 % Tween 20, 5 mM sodium azide) with 3 minutes of sonication between washes. The beads were finally dispersed at a concentration of 1 % (w/v) and stored at 4 °C.

I.9.2 Bead passivation

The beads were passivated by diluting six-fold in TE (Tris 20 mM pH = 8, 1 mM EDTA) and addition of β-casein to 1 mg/ml. The beads were vortexed for 10 minutes, washed once with TE, dispersed at a concentration of 0.2 % in TE and stored at 4 °C until the experiment.

I.9.3 DNA constructs for power spectra measurements

Measurement of power spectra used DNA constructs labeled with biotin and digoxigenin. The constructs were prepared by PCR using lambda DNA as the template, with biotinylated or digoxigenin labeled oligonucleotides (IDT) as primers. The constructs were used after standard PCR purification. To perform the experiment, 1 µl of 0.1 % anti-Dig coated beads were incubated with 1 µl of 10 nm DNA for 15 minutes at room temperature, followed by dilution in the experimental buffer (Tris 20 pH = 8, 130 mM KCl). Streptavidin-coated beads were used directly after passivation.

I.9.4 DNA constructs and proteins for transcription experiments

I.9.4.1 Plasmids and DNA templates

Plasmids pIA1127 (for expression of σ^{70}), pIA1234 (for expression of sortagged RNA polymerase), and pIA2-6 (used as a template for preparing DNA handles) were a gift from Irina Artsimovitch. Plasmid for the expression of sortase was a gift from David Liu.

A plasmid containing the T7A1 promoter followed by 8 repeats containing the His pause, originally described in reference [9], was modified by removal of the 1000 base sequence between the promoter and the repeats through digestion with AgeI and BamHI.

I.9.4.2 Preparation of DNA template and handles

To prepare the DNA template, the plasmids was restricted by BsaI-HF (8 units per µg DNA for 2 hours at 37 °C) and treated in parallel with shrimp alkaline phosphatase (0.4 unit per µg DNA) to generate a linear DNA with distinct, dephosphorylated 5′ overhangs. The enzymes were heat deactivated for 20 minutes at 65 °C, and the DNA was immediately treated with Klenow 3′-5′ exo− polymerase (1 unit per µg DNA) and 0.1 mM ddATP to generate an assisting force template. The reaction proceeded for 30 minutes at 37 °C, followed by heat inactivation for 20 minutes at 75 °C. The sample was then extracted 5 times with phenol-chloroform and once with chloroform, ethanol precipitated, and reconstituted in Tris 10 mM pH = 8, 0.1 mM EDTA. The purity of the DNA was assessed to be ∼ 88 % by agarose gel electrophoresis.

For producing 1.5 kb handles from the pIA2-6 plasmid, the following oligonucleotides were used:

```
For-biotin: 5' /5Biosg/GAAAGTCCGGCATCTCAATCCC 3'
Rev-BsaI:   5' ATGATACCGCGAGACCGATGTGGCTTCGGTCCCTTC 3'
```

Underlined bases correspond to sequences that anneal to the template plasmid.

The handle was prepared by PCR and cleaned by standard PCR cleanup, treated with BsaI-HF (5 units per µg DNA for 2 hours at 37 °C followed by heat inactivation for 20 minutes at 65 °C) and purified using PCR cleanup.

I.9.4.3 Preparation of σ^{70}

Plasmid pIA1127 was transformed into Rosetta2 bacteria. The bacteria were grown in 2 liters of 2YT medium supplemented with 1 % glucose, NPS (25 mM $(NH_4)_2SO_4$, 50 mM

KH_2PO_4, 50 mM Na_2HPO_4), 1 mM magnesium sulfate, 34 µg/ml chloramphenicol and 50 µg/ml kanamycin. The culture was grown at 37 °C to an OD_{600} of 0.5, transferred to 17 °C and IPTG was added to 0.1 mM. Induction proceeded for 16 hours.

For purification, the bacteria were dispersed in 80 ml of buffer A25 (Tris 20 mM pH = 7.5, 0.5 M NaCl, 10 % glycerol, 25 mM imidazole, 2 mM β-mercaptoethanol) supplemented with 0.1 mg/ml lysozyme and protease inhibitors (Roche). The bacteria were lysed by French press, and the lysate was clarified by centrifugation and filtration and loaded on a 5 ml Ni-NTA column. The column was washed with 20 ml buffer A25 and 20 ml A50 (A25 + 50 mM imidazole), and the his-tagged σ^{70} was eluted in A300 (A25 + 300 mM imidazole). TEV protease (prepared as described [99]) was added at a molar ratio of 1:40, and the reaction proceeded overnight at 4 °C while being dialyzed against A50. The sample was then passed again through an Ni-NTA column. The flow-through, containing non-his-tagged σ^{70} was collected, concentrated two-fold, and further purified by gel filtration on a sephacryl S300 column equilibrated with buffer B (Tris 20 mM pH = 7.5, 0.5 M NaCl, 10 % glycerol, 1 mM EDTA, 1 mM DTT). The protocol yielded ∼ 35 mg of pure σ^{70}. Aliquots were flash frozen in liquid nitrogen and stored at −80 °C.

I.9.4.4 Preparation of sortagged RNA polymerase holoenzyme

Plasmid pIA1234 was transformed into Rosetta2 bacteria. Sortag-RNAP was expressed using the same protocol as σ^{70}, except that ampicillin was used instead of kanamycin.

For purification we used a modified version of a published protocol [100]. The cells were dispersed in 75 ml of lysis buffer (Tris 50 mM pH = 6.9, 0.5 M NaCl, 5 % glycerol) supplemented by 0.1 mg/ml lysozyme and protease inhibitors, and lysed by French press. The lysate was centrifuged and filtered, and imidazole was added to 20 mM. The sample was loaded on a 5 ml Ni-NTA column. The column was washed with 30 ml of lysis buffer + 20 mM imidazole and the his-tagged RNAP core enzyme was eluted in lysis buffer + 250 mM imidazole.

To form the holoenzyme, the sample was incubated with a two-fold excess of purified σ^{70} overnight on ice. The sample was diluted ten-fold with buffer B0 (50 mM Tris pH = 6.9, 5 % glycerol, 0.5 mM EDTA, 1 mM DTT) and loaded on a heparin 5 ml column. To avoid overloading the column, the sample was divided into three portions that were loaded separately. A gradient of 50 mM to 1 M NaCl was used to elute the protein. RNAP holoenzyme was separated clearly from excess σ^{70}. The sample was dialyzed against buffer B50 (50 mM Tris pH = 6.9, 5 % glycerol, 50 mM NaCl, 0.5 mM EDTA, 1 mM DTT), and then purified further on a 1 ml monoQ column using a 50 mM to 1 M NaCl gradient (again, the sample was split into three portions loaded separately). Pure fractions were pooled, dialyzed against storage buffer (20 mM Tris pH = 7.5, 200 mM KCl, 0.2 mM EDTA, 0.2 mM DTT, 5 % glycerol), aliquoted, flash frozen and stored at −80 °C.

I.9.4.5 Biotinilation of sortag-RNAP

We obtained a peptide containing an N-terminal GGG tag with a biotin-modified lysine residue (Genscript): GGGGDGDY{Lys(biotin)}.

100 µl of 9.6 µM sortag-RNAP was reacted with a 200-fold excess of biotinilated peptide in 200 µl coupling buffer (Tris 50 mM pH = 7.5, 5 mM $CaCl_2$) containing 2 µM sortase (prepared as described [101]). The reaction proceeded for 60 minutes. At this point, imidazole was added to 25 mM and NaCl to 350 mM, and the sample was passed through 70 µl Ni-NTA beads to remove sortase and unreacted RNAP. The peptide was removed by dialysis into storage buffer, and the biotinilated RNAP was stored in storage buffer at −80 °C.

Part II

Pause sequences facilitate entry into long-lived pause states by reducing the forward transcription rate of RNA polymerase

This work is currently being submitted for publication. It has been reproduced here, with modifications, with the permission of all authors (Ronen Gabizon, Hanif Vahedian-Mohaved, Richard H. Ebright, and Carlos Bustamante).

This work was supported by the Howard Hughes Medical Institute (Carlos Bustamante), NIH Grants R01GM032543 and R01GM071552 (both to Carlos Bustamante), the US Department of Energy Office of Basic Energy Sciences Nanomachine Program under Contract DE-AC02-05CH11231 (to Carlos Bustamante), and NIH grant GM041376 (to Richard H. Ebright).

Transcription elongation by RNA polymerase (RNAP) is interspersed with sequence-dependent pausing. The processes by which RNAP enters paused states at sequence-dependent pause sites and the processes that stabilize these paused states have not been well characterized, due to the spatiotemporal limitations of methods previously used to study pausing. Here, by combining a high-resolution optical tweezers assay with improved data-analysis methods, we investigate the formation of paused states at enhanced spatiotemporal resolution and characterize their modulation by backtracking and by features in the nascent-RNA. We find that pause sites modify the dynamics of nearly all RNAP molecules in a force-independent manner, reducing their forward transcription rate, and not as previously thought, that they only affect the subset of molecules that enter long-lived paused states. We propose that the reduced rates at pause sites play a crucial role in pausing, by allowing time for the elongation complex to undergo conformational changes required to enter long-lived paused states. In addition, we find that pause stabilization by backtracking occurs in a stepwise fashion, with non-backtracked states or states backtracked by at most 1 base pair forming quickly, and further backtracking occurring slowly. Finally, we find that nascent-RNA features, such as RNA hairpins, act as modulators that enhance or attenuate pausing, depending on the sequence context.

Chapter II.1

Introduction

Transcription is a tightly regulated process in which RNA polymerase (RNAP) encodes the genetic information into RNA with either protein-encoding or structural and catalytic roles [102]. After initiating transcription from a promoter, RNAP enters the elongation phase, which consists of periods of processive nucleotide addition interspersed by pauses. Pausing plays critical roles in regulating transcription and in coordinating it with other processes that occur cotranscriptionally, including RNA folding [103], RNA processing and translation [10, 104]. The entry into paused states begins with the formation of an elemental paused state with inhibited transcription elongation [20, 55, 105]. In *E. coli*, pausing is known to occur at consensus pause sequence elements ($G_{-10}Y_{-1}G_{+1}$, where -1 corresponds to the position of the RNA 3′ end and $+1$ to the next nucleotide to be incorporated) [10, 11], through inhibition of the translocation step [11]. The paused states can be further stabilized by the formation of a nascent-RNA hairpin [17, 106] or by RNAP backtracking [17, 88, 107]. However, the events required to enter a paused state from active elongation are not well understood, because those events are beyond the spatiotemporal resolution of previous experiments. Optical tweezers experiments [9, 48, 55, 108–110] have been used to detect and characterize long-lived pauses (longer than 1 s), but have not been able to reliably and directly detect and characterize short, sub-second pauses. Since the time scale for processive nucleotide addition by RNAP at saturating nucleotide concentrations ranges from 25 ms to 100 ms per nucleotide [39, 111], a wide range of physiologically relevant time scales (from 25 ms to 1000 ms) has eluded direct study.

Here, we used a high-resolution optical tweezers assay with improved methods of data analysis to characterize transcription by *E. coli* RNAP through sequence-dependent pause sites at near-single-base-pair spatial resolution and an order of magnitude ($\sim$ 100 ms) improved temporal resolution. These improvements enabled us to answer three key questions about the mechanism of pause entry and the modulation of pauses by backtracking and nascent-RNA features. First, we find that strong sequence-dependent pause sites all involve a slowing of the forward elongation rate of the enzyme RNAP. This result supports a model in which

pause sequences reduce on-pathway elongation rates by RNAP, allowing it time to enter off-pathway reactions leading to long-lived paused states. Second, we find that stabilization of sequence-dependent pauses by backtracking involves two consecutive steps: a first step entailing rapid formation of a state that is either non-backtracked or backtracked by at most a single base pair, and a second step entailing slow conversion to a deeper backtracked and longer-lived state. Third, we find that nascent-RNA features, such as hairpins, can either increase or decrease the duration of sequence-dependent pauses, depending on the sequence context, most likely through interaction with the pretranslocated state of RNAP.

Chapter II.2

Characterization of transcriptional pausing at high spatio-temporal resolution

Previous optical tweezers studies of pausing relied on direct detection of pausing events by identifying time intervals in which the measured transcription velocity is below a defined threshold [48, 108]. These methods can consistently detect pausing events with lifetimes of at least $\sim$ 1000 ms, but must rely on extrapolation and/or other assumptions to infer the distributions of events occurring on shorter time scales [21]. The average pause-free velocity of RNAP at saturating concentrations of nucleotide triphosphates (NTP) is $\sim$10 bp/s to 40 bp/s, corresponding to a time scale for processive nucleotide addition of $\sim$25 ms to 100 ms [39, 108, 112], an order of magnitude shorter than directly accessible to previous methods ($\geq$ 1000 ms).

To overcome this limitation, we sought to fully characterize the dynamics of sequence-specific pausing, down to the $\sim$ 100 ms time scale. To this end, we developed a method that can accurately determine (1) the position of RNAP on the template sequence, (2) the time RNAP spends at each pause site for every crossing (the Pause Site Crossing Time), and (3) the pausing efficiencies at each pause site.

II.2.1 Repeat-based trace alignment

II.2.1.1 Data collection

In theory, since we have exact knowledge of the starting point of each trace, if we had an accurate measurement of the force on the tether and the worm-like chain parameters of the template under the experimental conditions, we would be able to accurately calculate the position of the polymerase on the template throughout the trace from the starting point. These conditions are not met for several reasons. First, there is a small uncertainty in the

starting position, since chasing the polymerase requires moving the beads in the chamber to a different location with a different buffer composition, which as mentioned before introduces small changes to the force (or trap distance under force feedback). Furthermore, the polymerase starts transcribing while the chamber is moving. This mixes the signal coming from transcription with signals coming from the viscous drag on the beads during the motion, which increases the uncertainty in starting position. Second, bead sizes and viscosity of the buffer are usually not precisely known, and therefore there is an uncertainty in the measured forces. This will change the conversion between nanometers and base pairs. In fact, the calibrations we later performed indicated that although our experiments were carried under nominally constant force, this conversion factor varies by 0.5 % to 1 % (figure II.4). Over a template 3000 base pair long, this may result in an error of 15 to 30 base pairs in positioning. Moreover, this error will not be distributed evenly across the trace, but will grow as the polymerase transcribes.

Instead, we modified a procedure first described by Herbert et al. [9]: we performed transcription experiments on a DNA template containing the T7A1 promoter followed by eight tandem repeats of a 239 bp sequence containing the his-leader pause site and four other known sequence-dependent pause sites [9, 55] (figure II.1). As we will show, the repeating pattern of sequence-dependent pausing that occurs throughout the repeats is sufficient to achieve accurate determination of the position of RNAP relative to the template sequence (to ± 3 bp).

Single transcription elongation complexes containing biotinylated *E. coli* RNAP halted at position 29 by NTP deprivation were tethered between 1 μm-diameter polystyrene beads held in high resolution optical traps (figure II.1; the instrument was described in part I). The experiment was performed in a laminar flow setup previously described [113]. The main channel of the chamber was formed by a flow coming from a reservoir containing buffer (HEPES 50 mM pH = 8, 130 mM KCl, 4 mM $MgCl_2$, 0.1 mM DTT, 0.1 mM EDTA, 20 μg/ml heparin and 10 mM NaN_3 (added as a singlet oxygen scavenger to reduce the extent of photodamage [86])) and a flow from a second reservoir containing a saturating NTP solution (1 mM UTP, 1 mM GTP, 0.5 mM ATP, 0.25 mM CTP [39]). The two flows form well separated regions in the chamber. Beads containing DNA handle were loaded on a side channel connected to the NTP side via a thin capillary, while beads containing the halted complex were loaded on a similar channel connected to the buffer side. Every experiment consisted of the following steps:

1. Trap a DNA handle beads in the NTP side;
2. Trap a halted complex bead in the buffer side;
3. Rub the beads against each other in the buffer side until a tether is formed (if at all);
4. If a tether is formed and it has the expected length, move the pair into the NTP side to restart transcription.

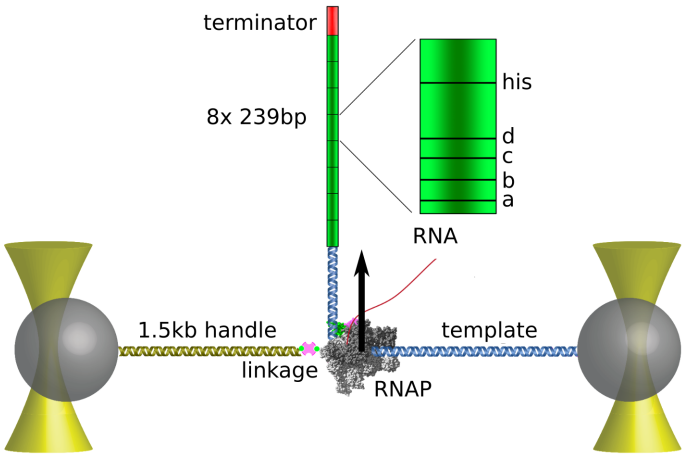


Figure II.1:
Experimental geometry. Biotinylated *E. coli* RNAP, halted on the template DNA, was tethered through a neutravidin bridge to a biotinylated 1.5 kb DNA ligated to the oligo-coated bead. The selection of which end of the template DNA to ligate to the other beads enables selection between assisting force and opposing force geometry. The major pause sites ('a', 'b', 'c', 'd', 'his') are indicated in the sequence of the repeat.

Subsequent elongation was monitored by measuring the extension of the tether (in nanometers) over time (figure II.2, left). Our experiments were performed using an active force clamp that moved one of the traps constantly to maintain the mean force constant. This maintained the force to within 0.1 pN within each trace, ensuring that a constant factor could be used to convert physical distances (in nanometers) to sequence positions (in base pairs). However, we discovered that small changes in calibration may arise due to the fact that the calibration and activity measurement were not performed in the same position in the chamber, possibly due to variations in the thickness of the glass or change in refractive index due to the different composition of the buffer in the NTP channel. Therefore, we performed an additional calibration in the NTP channel after the tether broke, and used this to calculate forces. Obviously, the feedback still had to be performed using calibration parameters measured in the buffer channel. This resulted in a small variation in measured force from tether to tether that rarely exceeded 0.5 pN. For each trace, data was collected until either the tether broke or the polymerase entered a pause longer than 200 s.

Prior to data analysis, we inspected the traces for irregularities. First, in a small subset of the traces, the breaking of the tether left a longer tether still connecting the beads, which indicates more than one molecule was tethered between the beads. These traces were discarded from further analysis. In 12 traces, large rips (20 nm to 100 nm in size) were observed at early

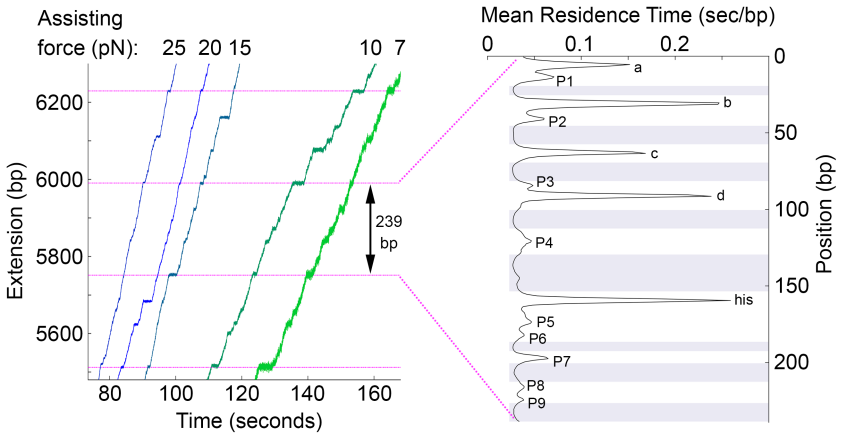


Figure II.2:
(Left) Representative traces obtained under assisting forces. Dashed magenta lines mark the locations of the 'his' pause sites in the template. (Right) The mean residence time histogram was calculated by averaging the time spent at each position in the repeat across all traces at all conditions, except RNase data, which was aligned separately (203 traces, table II.2.2). Pause sites are marked, and pause-free regions are shaded. For clarity, the mean of measurements up to the 95 % percentile is shown to remove to effect of rare pauses occurring outside the pause sites. However, data analysis was performed on the full dataset.

stages in the traces, followed by normal transcription and clean breaking at the end. In these traces, only data following the rip was used. Second, during the experiments, we determined that photodamage can be a significant effect in such systems, despite the addition of singlet oxygen scavengers. To test this, we performed experiments in which we held the beads in the traps for several minutes before attempting to obtain tethers. Tethers obtained this way were almost always inactive or exhibited extremely slow activity (at least threefold slower than average rates). We attribute this effect to the fact that before a tether is formed, the polymerase is frequently very close to the surface of the bead, where most reactive oxygen species are generated [86], while a tethered polymerase is far away from the surface and is thus better protected. When obtaining tethers less than one minute after trapping the bead, the majority of the tethers exhibited activity and tethers showing exceptionally slow activity typical of photodamage were identified easily and removed from analysis. The final dataset contained 432 transcription traces, out of which 251 traces reached the repeat region without entering a long-lived pause (> 200 s) or premature breaking of the tether.

We now describe the procedure used to achieve accurate determination of the position of RNAP.

II.2.1.2 Initial processing of data

The first step in the alignment procedure is the calculation of length in nanometers of the 239 bp repeat, which we term the *physical repeat length (PRL)*. This number is used to convert the physical distance transcribed by RNAP to the position on the template sequence. Since the traces were collected at a constant force, a constant PRL applies throughout each trace.

The repeats are located between 1107 bp and 3019 bp downstream of the starvation site. As an initial approximation, starting from an average rise per base pair of 0.33 nm, we thus estimated the position of the repeats as the region of the trace located between $1.02 \times 1107 \times 0.33\,\text{nm} = 363\,\text{nm}$ and $0.98 \times 3019 \times 0.33\,\text{nm} = 1005\,\text{nm}$ after the starting position; the factors of 1.02 and 0.98 are used to avoid including data from non-repeat regions. This approximation will be revisited later.

At 0.33 nm/bp, the PRL is 79 nm. Our algorithm requires the presence of at least two full repeats in order to align a trace; thus, we discarded any trace which covered less than $2 \times 82\,\text{nm} = 164\,\text{nm}$ of the repeats (where we used a slightly larger number for the period to provide some "buffer" to the analysis).

II.2.1.3 Residence time histogram

For each single-molecule trace, we first generated a residence-time histogram (RTH) by sorting the extension data into 0.1 nm bins; that is, we discretized the distance axis in small bins and asked, how much time did the trace spend in each of these windows? As the following discussion will make clear, the size of these bins must be chosen well below the expected size of a base pair; in order to work with round numbers, we set it to $(1/40)$nm, which is approximately $(1/13)$bp. Crucially, we find that, at least for the first step of the analysis (periodization), we do not need to downsample the data before computing the histogram; that is, the dataset at full bandwidth (800 Hz) is used to populate the histogram.

At regions of high transcription rates (as high as 40 bp/s), there will be 20 to 40 data points per base pair, which is 1 to 4 data points per bin. At such conditions the occasional empty bins are unavoidable. Such empty bins are problematic for the further analysis. Thus, we instead "connected" consecutive points in the trace (figure II.3) and populated histogram bins in proportion of what fraction of the connecting segments fell within the window. Such an approach also ensures that the residence time histogram changes continuously when the size or the origin of the bins in changed, rather than by discrete jumps whenever a data point crosses the edge between two bins.

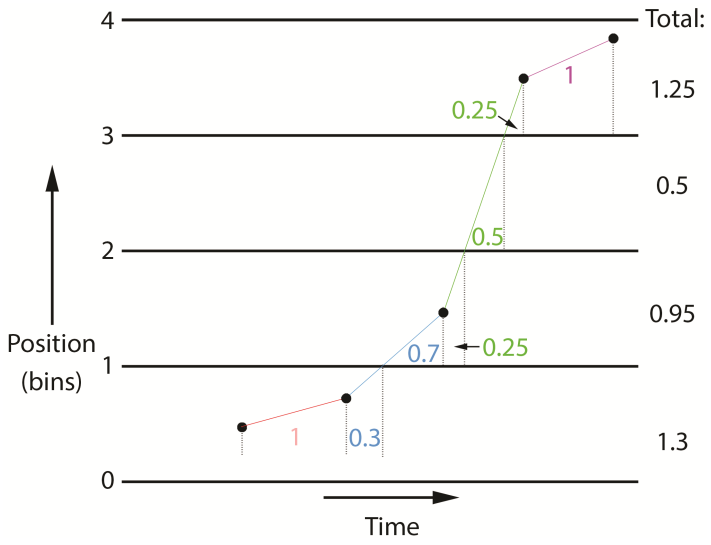


Figure II.3:
Illustration of residence time calculation. Residence times are calculated on segments between data points, according to which fraction of the segments is located in each bin. Therefore, even bins without data points (such as bin 2 to 3) will have a non-zero residence time.

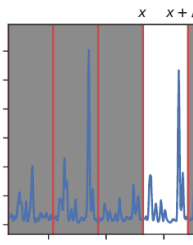
II.2.1.4 Periodization

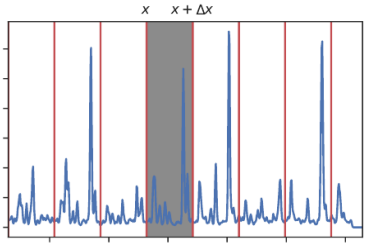
Having computed the residence time histogram (RTH) over the (expected) repeat region, we relied on the following observation in order to determine the correct nanometer-to-base pair conversion factor: the RTH over one repeat should look "similar" to the RTH over another repeat. Thus, we can try various candidate conversion factors (taken in a small range around the extensible worm-like chain prediction), corresponding to various values of PRL (for a 239 bp repeat, 75 nm to 81 nm, depending on force) and, for each of these conversion factors, quantify the "similarity" as suggested above. The PRL with the highest "similarity score" is then taken as the correct one.

Herbert et al. noted that such an evaluation of similarity can be carried out by "folding" the RTH over itself using the given PRL (that is, by summing the values in the RTH at positions $x, x + \mathrm{PRL}, \ldots$ for each x) [9]. A "good" period should yield very large spikes in the folded RTH due to the summing in phase of the pausing events over each of the repeats; in other words, some positions in the repeat should exhibit very long residence times. Thus, they computed the distribution of residence times in the folded RTH, and used the skewness (normalized third moment) of the log-residence-time distribution as the similarity score.

We found, instead, that a cross-validation style approach performs better to compute the similarity score. Specifically, we remove one repeat (the *testing set*), compute the folded histogram on the rest of the data (the *training set*), and compute the probability of observing the testing set under the hypothesis that the folded histogram gives the correct distribution. We then repeat this operation using each possible testing set, and use the average log-probability across all testing sets as the scoring criterion. The scheme can be graphically described as the calculation of

$$\max_{\Delta x} \int \log p \Bigg(\quad \Bigg| \quad \Bigg) \, \mathrm{d}x$$





Let us derive the expression for this average log-probability here. Let r_i be the residence time at position $i = 0, \ldots, N-1$ and P be the tentative period. We write $S_a^b = \sum_{i=a}^{b-1} r_i$ (sum of residence times from a to b), $S_{i[P]} = r_i + r_{i+P} + \ldots$ (sum of residence times at positions an integer number of periods from i), $S = \sum_i r_i$ (total residence time).

Consider the testing set corresponding to positions $i, \ldots, i+P-1$. The observed distribution from the testing set is $f_{i,j} = r_j / S_i^{i+P}$ (histogram counts at position j, divided by the total counts between i and $i+P$). The reference distribution for the training set is $g_{i,j} = (S_{j[P]} - r_j)/(S - S_i^{i+P})$,

where $i \leq j < i + P$ (folded counts at position a multiple of P away from j, excluding j itself, divided by total counts, excluding counts between i and $i + P$). The probability of observing the testing set given the training set is

$$L = \prod_{j=1}^{i+P-1} g_{i,j}^{r_j}; \tag{II.1}$$

and the log-likelihood for this testing set is thus

$$LL_i = \sum_{j=i}^{i+P-1} f_i(j) \log g_i(j) = \frac{1}{S_i^{i+P}} \sum_{j=i}^{i+P-1} r_j \log \frac{S_{j[P]} - r_j}{S - S_i^{i+P}} \tag{II.2}$$

(after normalization by $1/S_i^{i+P}$).

We now need to average these log-likelihoods over all possible testing sets, that is, $i = 0, \ldots, N - P - 1$:

$$\begin{aligned} LL &= \frac{1}{N-P} \sum_{i=0}^{N-P-1} LL_i \\ &= \frac{1}{N-P} \sum_{i=0}^{N-P-1} \frac{1}{S_i^{i+P}} \left[\sum_{j=i}^{i+P-1} r_j \log \left(S_{j[P]} - r_j \right) - r_j \log \left(S - S_i^{i+P} \right) \right] \\ &= \frac{1}{N-P} \sum_{i=0}^{N-P-1} \left[\frac{\sum_{j=i}^{i+P-1} r_j \log \left(S_{j[P]} - r_j \right)}{S_i^{i+P}} - \log \left(S - S_i^{i+P} \right) \right], \end{aligned} \tag{II.3}$$

where we have relied on the fact that $\sum_{j=i}^{i+P-1} r_j = S_i^{i+P}$.

The partial sums S_i^{i+P} and $\sum_{j=i}^{i+P-1} r_j \log \left(S_{j[P]} - r_j \right)$ can be computed efficiently by first computing the cumulative sums starting from index 0 and then taking the difference between the cumulative sums at the two endpoints of the sum.

While this approach works well for finding the correct period size of traces that exhibit moderately strong periodic pausing, we found that for traces where the pausing is very weak, the trace with the highest similarity score tends to pick a period that is "too small" (by comparison with other traces collected at the same force). We believe that this is due to the fact that the computation of similarity scores for different periods (expressed in terms of physical size) compares the RTH over different numbers of bins. For example, consider the case of a perfectly uniform RTH with P bins (each with a relative occupancy of $1/P$). The

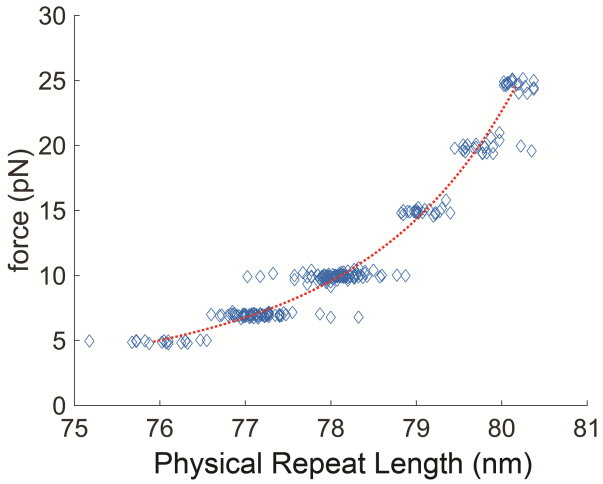


Figure II.4:
Repeat lengths vs. force for all aligned traces ($N = 251$). Worm-like chain model fitting was performed on the average lengths obtained at every force ($R^2 = 0.999$).

log-likelihood for any training set would be

$$LL_i = \sum_{j=i}^{i+P-1} \frac{1}{P} \log \frac{1}{P} = -\log P\,. \tag{II.4}$$

Thus, we need to correct the similarity score by taking $-\log P$ as the base score (which effectively penalizes small periods): the final similarity score that we maximized in order to find the physical period size was $LL + \log P$.

The period sizes are expected to increase with force as the DNA becomes more extended. This is observed when plotting the period sizes versus the force (figure II.4). Fitting the mean period sizes obtained at each force to an extensible worm-like chain model yielded a persistence length of (30 ± 18) nm; the stretch modulus could not be fitted with high accuracy. The value of the persistence length is consistent with established parameters for DNA.

II.2.1.5 Inter-trace alignment

Having computed the correct period size, or nanometer-to-base pair conversion factor, for each trace, we next needed to find the relative position offset between each trace (all distances being now expressed in base pairs). For this purpose, we relied on a similar cross-validation strategy: we shifted the RTH of each trace upstream or downstream until they matched each other as well as possible, as measured by the average log-probability of observing a given

RTH (testing set) if the underlying probability distribution is given by the average of the other RTHs (training set).

Note that, in order to align various traces with each other, the bin size of the RTHs need to be identical; a new RTH was thus computed for each trace with a bin size of $(1/10)$bp (using the same procedure as above). Additionally, we found that for this purpose, a better quality alignment was obtained by downsampling the traces before computing the RTH; the traces were downsampled to 5 Hz.

In this case, the expression for the average log-probability is simpler. We define $r_{u,i}$ the average residence time of the uth trace at position i, $S_u = \sum_j r_{u,i}$ (total average time for trace u), $S_i = \sum_u r_{u,i}$ (total time at position i) (note the abuse of notation), and $S = \sum_{u,i} r_{u,i}$. The log-probability associated with the uth trace is (similarly to above)

$$LL_u = \sum_i r_{u,i} \log \frac{S_j - r_{u,i}}{S - S_u} \tag{II.5}$$

and the average log-probability across training sets is

$$\begin{aligned} LL &= \sum_u LL_u \\ &= \sum_{u,i} r_{u,i} \left[\log \left(S_j - r_{u,i} \right) - (S - S_u) \right] \\ &= \sum_{u,i} r_{u,i} \log \left(S_i - r_{u,i} \right) - \sum_u S_u \log \left(S - S_u \right) \ . \end{aligned} \tag{II.6}$$

The second term is independent of the relative offsets between the tethers and can be ignored.

Unfortunately, due to the large number of parameters in this maximization problem (one offset for each trace, except the first trace which can be taken as an (arbitrary) reference), and the presence of multiple local maxima, we found that the global optimizer on which we relied (`scipy.optimize.basinhopping` [114]) is unable to find a satisfactory solution.

We thus relied on a simpler approach first: we computed, for each pair of average residence time histograms $(r_{u,i})_i$, $(r_{v,i})_i$ the correlation between the two traces ($(r_u * r_v)_i = \sum_j r_{u,j} r_{v,i+j}$), and found the relative shift $\delta_{u,v}$ that maximized this correlation. A natural idea is then to find shifts δ_u for each individual trace such that $\delta_u \ominus \delta_v = \delta_{u,v}$ for all u, v; where $\ominus$ denotes the difference taken modulo the number of bins. However, this system is overdetermined and has no solution; additionally, the presence of the modulo term makes classic linear algebra techniques, such as least squares, not directly applicable.

It remains natural to attempt to find an approximate solution for the system $\delta_u \ominus \delta_v = \delta_{u,v}$. One could, for example, attempt to find the minimizer

$$\arg\min_{(\delta_u)} \sum_{u,v} \left((\delta_u \ominus \delta_v) \ominus \delta_{u,v} \right)^2 \tag{II.7}$$

using a global optimizer. However, as it turns out, not all $\delta_{u,v}$ are "reliable": for some pairs of traces, the maximum in the correlation matches two "incorrect" peaks. Thus, we need to use a robust alternative to least squares, which is able to ignore an unsatisfiable equation as long as most others are properly handled. A classical example is least absolute deviations,

$$\arg\min_{(\delta_u)} \sum_{u,v} \left| (\delta_u \ominus \delta_v) \ominus \delta_{u,v} \right| . \tag{II.8}$$

An additional correction is needed to make the minimization problem better behaved. A problem of the least absolute deviations formulation is that it is not strictly convex, even locally; specifically, if, say, $\delta_u \ominus \delta_v < \delta_{u,v}$ and $\delta_u \ominus \delta_w > \delta_{u,w}$, then the effect of slightly changing δ_u on these two terms will cancel out each other; that is, the target function is degenerate. We fix this issue by rendering the target function locally strictly convex,

$$\arg\min_{(\delta_u)} \sum_{u,v} \sqrt{\left((\delta_u \ominus \delta_v) \ominus \delta_{u,v} \right)^2 + \epsilon^2} \tag{II.9}$$

where ϵ is a small regularizer, chosen to be equal to one base pair.

It turns out that this target function is well-behaved enough so that a generic global minimizer (`scipy.optimize.basinhopping`) can find a reasonable minimum for it quickly. This minimum is then used as an initial point to the log-probability maximization problem we formulated in the first place.

Having found the set of offsets that maximize this probability, each trace may still be "off" their correct position by an integer number of periods. This error is fixed by shifting each trace by an integer number of repeats so that their start position (which, as explained above, is relatively poorly defined due to the need to move from the buffer channel to the NTP channel) is no more than half a repeat away from the "expected" initial tether length.

In an initial alignment, we set the positions of the major pauses 'a', 'b', 'c', 'd', and 'his' within the 239 bp repeat to be 5, 30, 63, 90 and 158, respectively, thus maintaining the distance between the pause sites in the sequence. We shifted the complete, aligned residence time histogram to maximize the time spent at those sites (figure II.5). First, we observed that site 'c' was clearly shifted $\sim$ 1 bp upstream relative to the expected position. To validate the positions of the pause 'c', we performed a bulk transcription experiment covering the region of pauses 'c' and 'b' (figure II.6). The results show a strong pause site 1 bp upstream compared to the previously known position of pause 'c', but with a weaker pause following

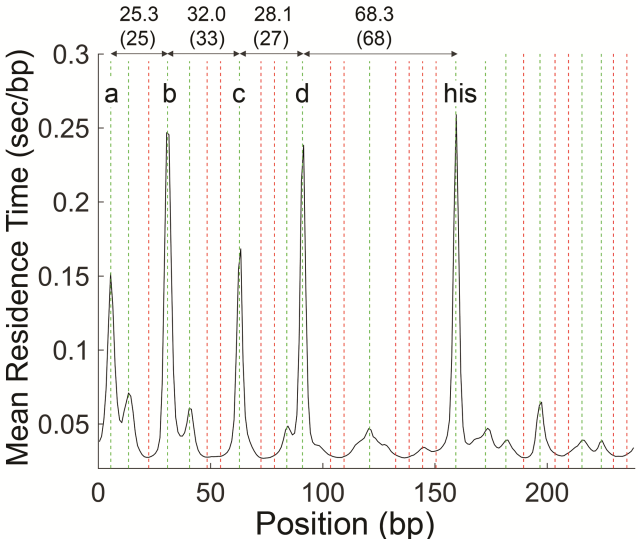


Figure II.5:
Residence time histogram of data collected at all conditions (excluding RNase samples). Calculated distance between the main peaks are shown above the arrows (top value: calculated value; value in parenthesis: expected according to Herbert et al.). The colored vertical lines indicate the positions used for residence time calculations. Pause site positions are shown in green and major sites are labeled. Reference sites at the pause-free region (red) were analyzed and the residence times calculated in them were aggregated for each condition.

it. The same result was observed when using the NTP concentrations used in earlier studies (data not shown) [9]. The source of the discrepancy is thus unclear. Regardless, we moved the position of pause 'c' accordingly. It should be also noted that the observed distance between those sites was not an integer number of base pairs. This may be due to several reasons: a possible difference in translocation state between the pause sites, a difference in the rise per base pair along the template [9], as well as inherent limitations on the accuracy of our data and alignment. We used the detected pause positions at this stage as the expected positions for pause scoring described later.

Remember that at the beginning of our alignment procedure, we needed to assume what section of the trace corresponded to the repeat region, based on an approximate estimate of the size of a base-pair and of the position of the starvation site. As the output of our alignment procedure consists of more accurate estimates of both quantities, we used these outputs to better estimate the position of the repeat region, and repeat the entire alignment procedure (only for one iteration).

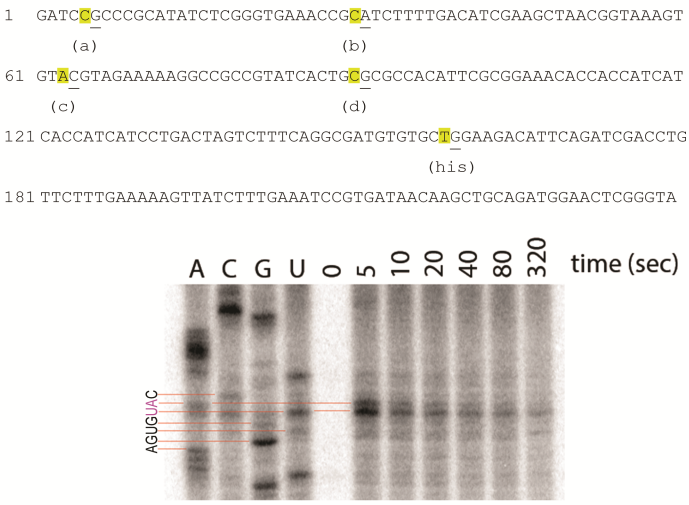


Figure II.6:

Top: sequence of the repeat with the expected pause sites according to Herbert et al. The underlined base denotes the +1 base on the non-template strand, while the yellow color indicates the 3′ end of the RNA at the pause site. Bottom: bulk transcription experiment surrounding the area containing pause 'c', under the same buffer conditions and NTP concentrations used in the trapping assay. An RNA ladder was generated by performing the reaction in the presence of 3′-deoxy-NTPs. RNAP pauses strongly after incorporating U, and more weakly after incorporating the subsequent A, indicating the main pause site is 1 bp upstream of the expected position.

II.2.1.6 Results of the alignment

The resulting sequence-dependent pausing profile is presented in figure II.2 (right). We detected the strong pause sites characterized by Herbert et al. ('his', 'a', 'b', 'c', and 'd'), nine other sequence-dependent pause sites with shorter residence times (labeled P1–P9), and the almost entirely pause-free regions between the pause sites. The weak pause sites P1–P9 were partially evident in previous studies [9] but with lower resolution. They appear as peaks across all tested conditions and forces, and display the high force sensitivity characteristic of pause site, is contrast to non-pause sites (as will be shown later, figure II.20). These results indicate that P1–P9 are weak pause sites and not random fluctuations in transcription rates. Table II.2.1 contains the sequences of the identified pause sites, and table II.2.2 contains the average transcription rates in pause-free regions. The 'his', 'a', and 'd' pause sites exhibit three out of three matches to the consensus pause element, $G_{-10}Y_{-1}G_{+1}$ [10, 11], whereas the 'b' and 'c' pause sites exhibit only two matches. In contrast, the weak pause sites, P1–P9, exhibit at most one out of three matches to the consensus pause element. We also checked if the weak pause sites match the more extensive consensus sequence $G_{-11}G_{-10}T_{-3}G_{-2}Y_{-1}G_{+1}$ [10, 12], and found only one site (P9) that displayed more than one match. We focused our study on the five strong pause sites, as well as the site 'P2', which displayed high sensitivity to RNase, as described later.

Table II.2.1:
Sequence of identified pauses sites. Next-to-last base is the 3′ end of the RNA (for sites 'a', 'b', 'c', 'd' and 'his': known from bulk assays and single molecule data; for sites P1–P9: estimated from single molecule data). Underlined bases indicate bases from the consensus motif $G_{-10}Y_{-1}G_{+1}$.

site	sequence (−11 to +1)	consensus bases
a	CGGGTAGATCCG	3/3
b	GGTGAAACCGCA	2/3
c	CGGTAAAGTGTA	2/3
d	CGTATCACTGCG	3/3
his	CGATGTGTGCTG	3/3
P1	TCCGCCCGCATA	1/3
P2	CATCTTTTGACA	1/3
P3	AGGCCGCCGTAT	1/3
P4	ACCACCATCATC	1/3
P5	AAGACATTCAGA	0/3
P6	AGATCGACCTGT	1/3
P7	TTGAAAAAGTTA	1/3
P8	AATCCGTGATAA	0/3
P9	TAACAAGCTGCA	1/3

Table II.2.2:
Number of traces, number of analyzed repeats, and pause-free velocities under all tested conditions. Errors are 95 % confidence intervals estimated by bootstrapping.

condition	# traces	# analyzed repeats	velocity (bp/s)
+25 pN	18	152	42.7 ± 0.3
+20 pN	22	184	37.6 ± 0.3
+15 pN	23	195	33.5 ± 0.3
+10 pN	17	139	34.4 ± 0.2
+10 pN + GreB	16	131	26.1 ± 0.2
+10 pN + RNase	26	218	32.0 ± 0.3
+7 pN	17	144	28.4 ± 0.3
−5 pN	17	142	26.2 ± 0.2
−7 pN	18	153	28.3 ± 0.2
−7 pN + GreB	17	148	19.4 ± 0.2
−7 pN + RNase	19	159	24.6 ± 0.3
−10 pN	16	119	22.9 ± 0.2
−10 pN + GreB	17	144	18.9 ± 0.2

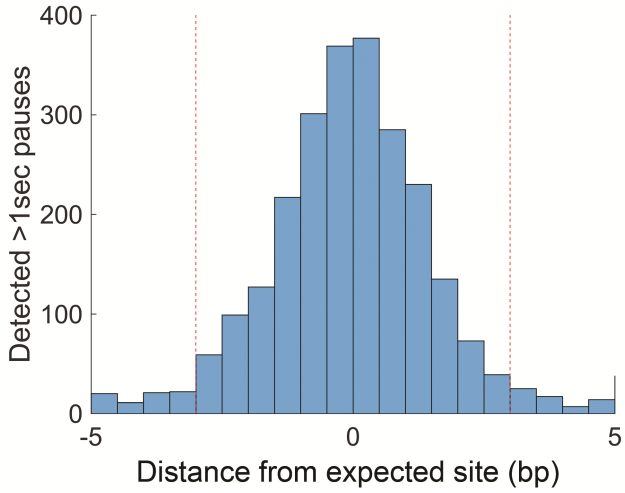


Figure II.7:
Distribution of positions of detected > 1 s pauses relative to the sequence positions of the major sites 'a', 'b', 'c', 'd' and 'his'. Red dashed lines indicates the ± 3 bp range.

II.2.2 Extraction of pause lifetimes

The second requirement—accurate determination of the pause site crossing times—was fulfilled as follows. We estimated that the location of RNAP at any given position in the data was known to within ± 3 bp (figure II.7). Given this localization accuracy, we could draw, around each expected pause site, a 6 bp window in which the actual pause site must be located. Since no pause sites were found within 6 bp of one another, each window surrounding a pause site contains $\sim$ 6 steps, one of which corresponds to the crossing of the pause site itself and the others to the crossing of pause-free sites. Next, we made the key assumption that within each of these 6 bp windows, the position of the pause site corresponds to the slowest step. To estimate the crossing time of the pause site, we de-noised the traces using total variation regularization (section II.2.2.1, figures II.8 and II.9), searched within each 6 bp window for the 1 bp step that took the longest to cross, and took the duration of that slowest 1 bp as the crossing time at that pause site (section II.2.2.2).

II.2.2.1 Trace regularization

While linear filtering of the data is commonly used as a preprocessing step to detect pause events in an optical tweezers trace, such a method is known to exhibit low sensitivity to fast, sub-second pausing events. Instead, we relied on total variation regularization to compute

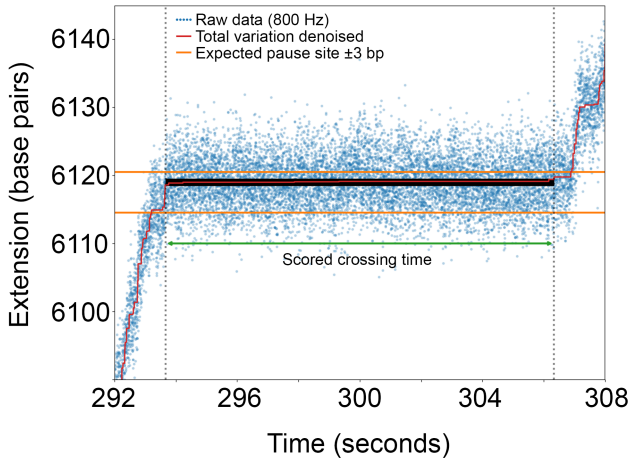


Figure II.8:
Total variation denoising and computation of pause site crossing times. The total variation denoising (red) of the raw data (blue) consists of flat segments separated by discrete jumps. One of these segments occurs in the vicinity of an expected pause site. 1 bp windows are drawn in the ± 3 bp range surrounding the expected pause site; the window that took the longest to cross (solid black) is used to define the pause site crossing time for the crossing of this pause site.

pause lifetimes [52]. A side-product of this regularization is the ability to detect rips and zips.

Let us consider a trace $y = (y_0, y_1, ...)$. For traces where no backtracking occurs, a natural way to regularize the trace is *isotonic regression*; namely, finding the trace $\hat{y}$ that is non-decreasing ($y_i \leq y_{i+1}$ for all i) and that minimizes the sum of squared errors, $\sum_i (y_i - \hat{y}_i)^2$.

However, backtracking is clearly observed, if somewhat rarely, in our dataset. In order to allow our fit to go backwards whenever needed, we write a target function that penalizes both deviation from the measurements (sum of squared errors) and "excessive" following of the noise spikes:

$$\arg\min_{\hat{y}} \left[\sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_i |\hat{y}_{i+1} - \hat{y}_i| \right]. \tag{II.10}$$

For example, note that the second term is equal to the end to end distance of $\hat{y}$ if $\hat{y}$ never moves backwards; each backtrack increases the term by twice the backtrack depth. The factor λ indicates the relative importance we give to the two terms.

Efficient algorithms exist to find the minimizer $\hat{y}$; we relied on the implementation of Johnson's dynamic programming algorithm [53] in the `prox_tv` package.

The remaining question is the choice of the relative weight, λ, between the fidelity term (mean square error) $\sum_i (y_i - \hat{y}_i)^2$ and the regularization term (total variation) $\sum_i |\hat{y}_{i+1} - \hat{y}_i|$. We relied on the L-curve method to pick such a value [115]. The L-curve is the parametric curve obtained by plotting the points

$$\left(\left(\sum_i (y_i - \hat{y}_i)^2 \right)^{1/2}, \sum_i |\hat{y}_{i+1} - \hat{y}_i| \right) \tag{II.11}$$

for various values of λ in log-log scale (we plot the *root* mean square error so that both terms have the dimension of a distance). When λ is small, increases in λ tend to greatly decrease (i.e., improve) the regularization term (by avoiding to follow the large number of fast spikes due to noise), while only minimally increasing (i.e., worsen) the fidelity term (as the smoothed out peaks are small in amplitude). Conversely, when λ is large, increases in λ tend to only minimally decrease the regularization term (because most of the spikes have been smoothed out), while greatly increasing the fidelity term (because all that is left to do is to smooth out the large, "true" movement in the trace). The presence of these two regimes gives the L-curve an "L" shape, where the two branches correspond to the two regimes described (figure II.9). The "corner" of the L corresponds to a choice of λ where the compromise between the fidelity term and regularization term is, in a sense, "optimal".

The end result of this procedure is a trace that consists of exactly flat regions, separated by sharp jumps ("staircasing effect" [52]) (figure II.8). Jumps of a size greater than 4 nm were

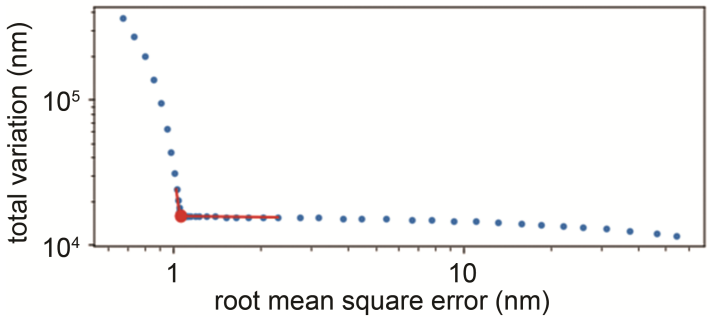


Figure II.9:
L-curve for selecting the value of the regularizer for the 10 pN assisting force dataset. Individual blue dots correspond to individual values of λ (taken at a constant ratio, $2^{1/3}$, of each other), with greater values of λ at the bottom right of the curve. The "vertical" branch of the curve corresponds to the regime where increases in λ mostly decrease the total variation term; the "horizontal" branch to the regime where increases in λ mostly increase the root mean square error. The two red segments are tangents to the L-curve at its inflection points (where the L-curve switches from "curving clockwise" to "curving counterclockwise" and vice-versa); their intersection is marked by a red dot; the blue dot closest to the red dot defines the chosen value of λ.

considered as rips or zips in the raw data; whenever one was detected, the trace was split into two sections (before and after the jump) that were analyzed separately.

II.2.2.2 Calculation of pause site crossing times

The regularized trace was then used for scoring the pauses. Again, remember that the result of the regularization is a trace that is nearly nondecreasing, with backtracks occurring only at a few, discrete positions.

We first converted the trace to a fully non-decreasing one by replacing its value during any backtrack by the maximum value attained so far. Physically, this operation can be understood as tracking the nature of the 3' end of the RNA, rather than the position of the polymerase itself.

We could then compute the pause site crossing time as follows: for each site, we defined a fixed range surrounding the pause site symmetrically, with the range wide enough to ensure that the pause site will be located within the tested range, but narrow enough to exclude other pause sites. Within this range, the polymerase takes steps through multiple pause-free sites and a single pause site. To calculate the crossing time of the pause site, we assumed that the crossing of the pause site was the slowest step, and therefore picked, among all the

1 bp-wide windows that one can draw within the tested range, the window that took the longest to cross (figure II.8).

For this purpose, we first needed to assess how large this range needs to be to capture most of the pausing events occurring at any site, which is, essentially, a measure of the accuracy of the positioning individual RNAP molecules. The P1 site is located ~ 8 bp downstream of 'a', and P2 is located ~ 10 bp downstream of 'b', and they are very clearly resolved. Therefore we set the range of 5 bp on each side as maximum range necessary. To find if we can reduce this range, we applied the pause detection algorithm used previously [9] with slight modifications. The data was downsampled to 200 Hz and filtered with a 3rd order Savitzky-Golay filter with a 1 s window. The instantaneous velocity was computed at every point and a velocity histogram was calculated using 0.2 bp/s bins for all the traces collected at each condition. The region from -2.5 bp/s to 2.5 bp/sec was then fit to a Gaussian. We assumed that all rates lower than the center of the Gaussian are associated with pausing events and that the distribution of rates for pause sites is symmetric. Using this assumption we calculated the histogram for rates associated with pausing above the center. This enabled us to calculate, for every possible rate, what is the probability that this rate is associated with a pause. We defined a threshold rate as the rate below which this probability is at least 90 %. We used this threshold to score pauses, while combining any adjacent pauses that were spaced < 1 bp apart.

After pause scoring, for each scored pause we found the closest major site. Then, for the ± 5 bp range around every major site we plotted a histogram of the distance of scored pauses from the site (figure II.7). We found that a range of 3 bp at every direction contains ~ 94 % of the observed pausing events. We therefore chose a range of ± 3 bp for pause scoring. Crossing time distributions obtained using the two methods showed good agreement for pauses longer than 1 s and sometimes even shorter pauses (figure II.10). However, our new scoring method finds a crossing time for every crossing of the pause site, therefore enabling us to characterize the behavior at very short time scales, as discussed below.

II.2.2.3 Resolution of the crossing time distributions

When performing this calculation, we essentially measure the slowest step between six steps taking place in the analyzed window (given our localization accuracy of ± 3 bp). In pause-free regions, this will naturally result in crossing times that are longer than the average pause-free dwell. Crossing time distributions in pause-free sites can be approximated by a distribution of the maximum of six identical exponentials, yielding transcription rates close to the measured average values (figure II.11). The behavior at long time scales at opposing forces is not well fit by this model, possibly reflecting the higher tendency to enter pauses outside the pause regions.

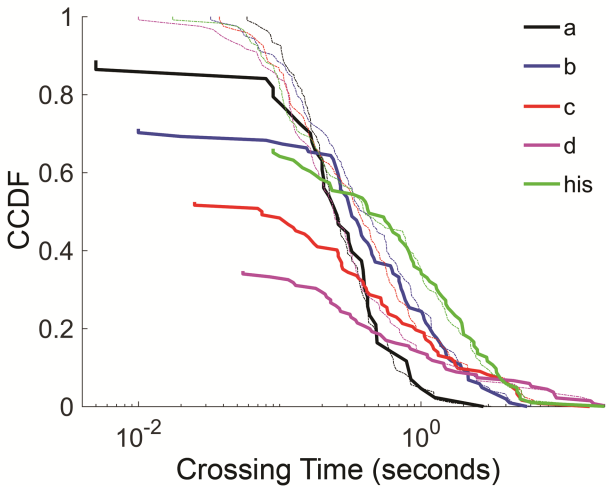


Figure II.10:
Comparison of pause site crossing time distributions calculated from previously published algorithms (solid lines) to crossing time distributions calculated using the method presented in this manuscript (dashed lines), for 10 pN assisting force. The distributions in solid lines were rescaled to equalize the value of CCDF at 1 s and illustrate the difference arising at shorter time scales.

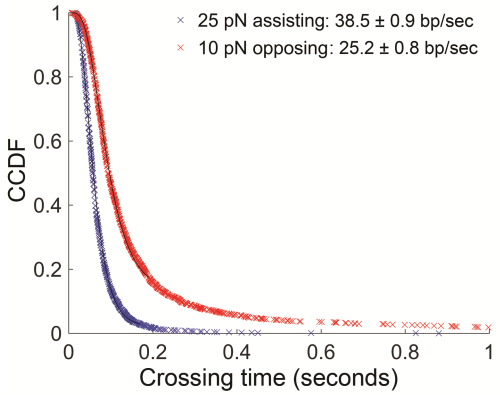


Figure II.11:
Crossing time distributions calculated at 16 pause-free sites and aggregated, at 25 pN assisting force and 10 pN opposing force. The data was fitted to a maximum of six exponentials with the same rate. We fit the data using successively truncated datasets (by removing long events) and selected the fit with the best Kolmogorov-Smirnov statistic when compared to the experimental data (fitting ~ 98 % of the 25 pN assisting force data and 80 % of the 10 pN opposing force data).

At pause sites, the behavior of the calculated crossing times is more complex. If the polymerase happened to have crossed the pause site relatively quickly (for example, because it did not pause or because RNAP was in the paused state for a relatively short time), there is a significant probability that the event captured by our method will be the crossing of one of the pause-free sites surrounding the pause site, and therefore some crossing times measured at short time scales will come from crossings of pause-free sites. As the pause site crossing becomes longer, the probability of capturing increases. To estimate the probability of capturing a crossing of the pause site, we assumed the five surrounding pause-free steps are exponentially distributed and calculated the probability that a pause-site crossing of a given length will be longer than the other five steps. The calculation shows that pause-site crossings as short as 4.5 times longer than the average pause-free dwell (equivalent to 125 ms to 250 ms) will be captured with $> 95\,\%$ probability. A significant fraction of shorter (50 ms to 100 ms) crossings will also be captured. Indeed, as shown in the main text, the crossing time distribution at pause sites can be distinguished from the distribution at pause-free sites even at very short (< 200 ms) time scales.

II.2.2.4 Results of pause lifetime extraction

We measured the crossing time distribution (CTD) at each pause site, as well as in pause-free regions, for which the data were aggregated into a single distribution (figure II.12 and figure II.11). The heterogeneity within the reference sites is small compared to the difference between reference sites and pause sites (both P1–P9 and 'a', 'b', 'c', 'd', 'his'), justifying the aggregation of data from all reference sites.

A key feature of the method we use for calculating crossing times is that no element of pause detection is employed. Instead, the method calculates the crossing time for every instance of the tested site, whether a pause occurred or not, yielding full distributions of all crossings of the pause sites. The CTDs at pause sites can be clearly distinguished from the distribution measured at non-pause sites under the same conditions using the same method (which we term the reference CTD), down to a time scale of 100 ms (figure II.12).

II.2.3 Estimation of pausing efficiencies

The third requirement—accurate determination of pausing efficiencies—was fulfilled by employing a non-parametric computational approach, as follows. For each pause site, we define the pausing efficiency as the probability that an RNAP molecule will reduce its transcription rate when crossing the site. In previous studies, the number of pauses shorter than 1 s was estimated by extrapolation of the pause lifetime distribution measured at ≥ 1 s time scales to shorter times. To test this method, we tried to fit the 1 s to 20 s region in the

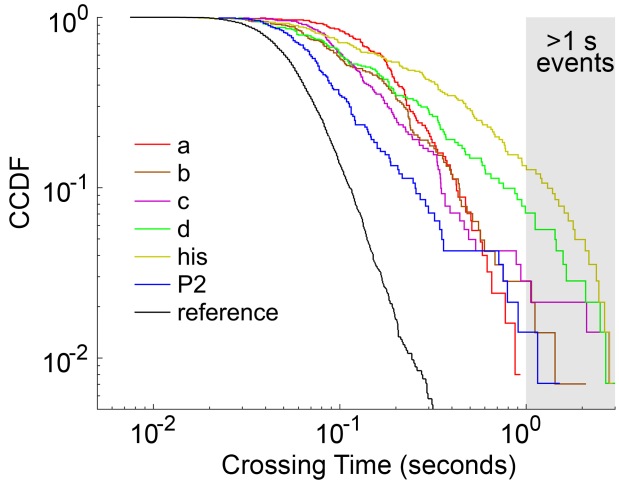


Figure II.12:
Crossing time distributions for different sites measured at 25 pN assisting force. The complementary cumulative distribution function (fraction of events longer than a given crossing time, CCDF) is plotted. The gray shaded area marks the time scales accessible to previous experiments.

crossing times to an exponential distribution (figure II.13). We encountered several issues: first, an exponential distribution did not give good fits consistently for any site. Second, in most sites the crossing time log-cumulative distribution becomes steeper at the 0.5 s to 1 s time scale compared to longer time scales, indicating faster dynamics dominating at this time scale; that is, there are more 0.5 s to 1 s events than would be expected from an exponential extrapolation from longer time scales. In other words, extrapolation to $t = 0$ from events longer than 1 s inherently underestimates the pausing efficiency, with the effect probably being the greatest at high assisting forces, where very few of the events are longer than 1 s. Finally, at weak pause sites, or at high assisting forces, the number of events longer than 0.5 s to 1 s was frequently very low, making such fits very unreliable.

In contrast, we have directly measured the full pause site crossing time distributions. Due to the stochastic behavior of single RNAP molecules, at the 50 ms to 100 ms time scale, events cannot be unambiguously assigned as pause-free or paused, contrary to very fast (< 50 ms) or very slow (> 1 s) events, that can be assigned with certainty as pause-free or paused, respectively (figure II.14). This inherent limitation is independent of the resolution of the method. We therefore calculated the pausing efficiency at each site by comparing the CTD measured there to the reference CTD (measured at non-pause sites), modifying an analysis of power law tails presented by Clauset et al. [54].

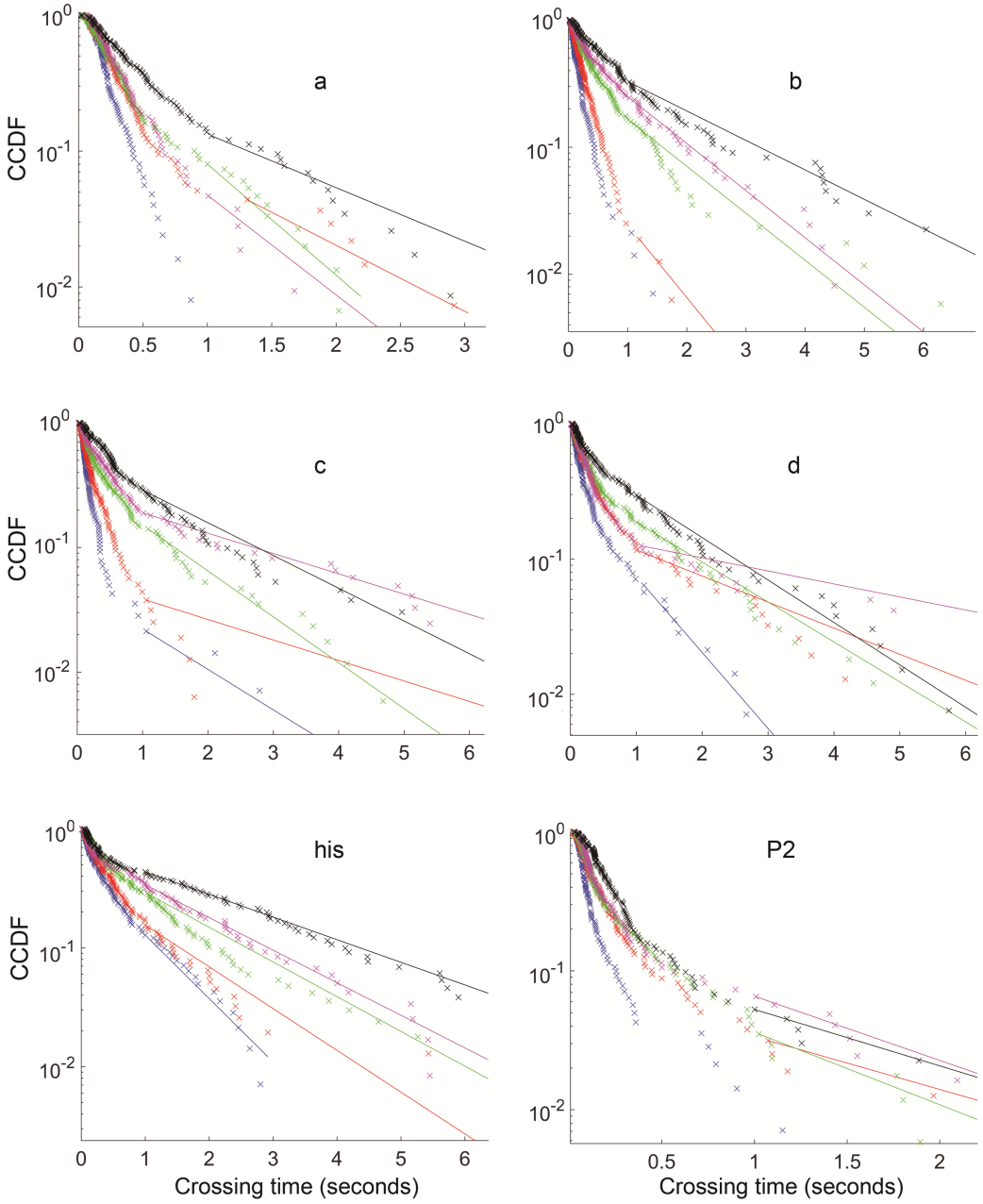


Figure II.13:
Exponential fits to crossing times in the 1 s to 20 s range for the major pause sites and P2 at five assisting forces. For pauses 'a' and P2 at 25 pN, the fit did not converge.

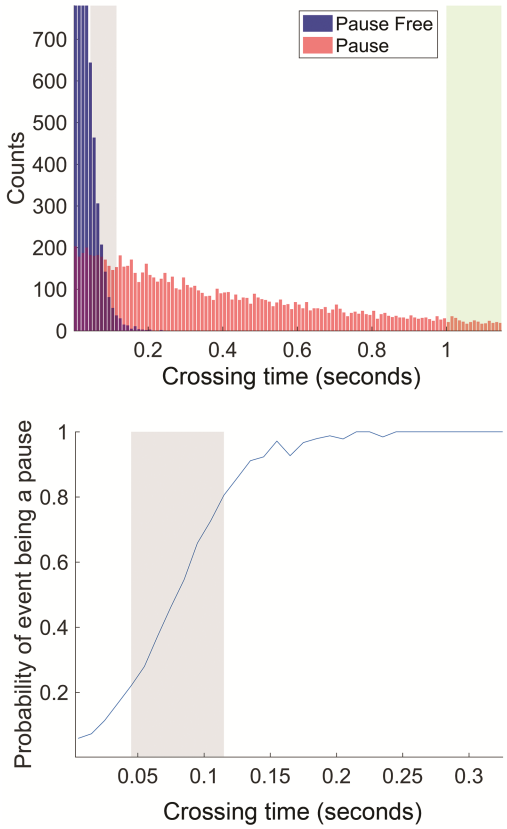


Figure II.14:
(Top) Illustration of the overlap between the distribution of crossing times for paused states and pause-free states. For each curve, 10,000 points were randomly generated from an exponential distribution with a mean of 0.025 s (blue, pause-free), or from an exponential distribution with a mean of 0.5 s (red, pause). It is very clear that events at the time scale of 1 s or longer (green shaded area) can only be pauses; with higher temporal resolution, a temporal regime in which an event has similar probabilities of being derived from a paused states or a pause-free state is reached (gray shaded area). Pausing efficiency is estimated by how much the distribution is enriched with longer events relative to the distribution measured at pause-free sites. Exponential distributions were selected for the simulation for simplicity—the principle would apply regardless of the shape of the distribution of pause crossing times. (Bottom) The probability that an event with a particular duration arises from the paused state was calculated for the case of 50 % pausing efficiency (half of the events arise from the paused state). The probability increases with the duration of the event.

We assume that for a certain fraction of the observed crossing times, RNAP did not enter a paused state, and that the crossing times of these pause-free events have the same distribution as the crossing times measured at the pause-free sites (the *reference* distribution). The remaining events arise from the paused state, and their fraction is the pausing efficiency. As explained in the main text, long events arise from the paused state with very high probability, while events in the 50 ms to 150 ms range have similar probabilities of arising from a pause or from pause-free dynamics (figure II.14). The probability that a particular event arises from a pause increases gradually as the crossing time increases.

Let us consider a distribution of crossing times at the pause sites (figure II.15, blue curve), and a distribution of crossing times at the reference sites (figure II.15, black curve). Imagine that we were only able to observe, both at pause and at reference sites, the events that lasted no longer than a cutoff time τ (figure II.15, gray vertical line)—and were oblivious even to the existence of longer events. We could test how well the pause and reference distributions match each other by computing a distance statistic between the distributions (in our case, the Kolmogorov-Smirnov statistic). We compute this statistic for all possible τs, and pick the one for which the two distributions are the closest to each other as the "cutoff time" (typically, if τ is chosen too small, then statistical fluctuations due to small sample size lead to a larger distance between the truncated distributions; if τ is chosen too large, it is the intrinsic difference between the underlying distributions that leads to a large distance between the truncated distributions) (figure II.15).

The similarity of CTDs below the cutoff at the pause site and at pause free sites suggests that events shorter than the cutoff arise from the same (pause-free) state in both cases (figure II.15, light green vertical bar). As stated above, there are also pause-free events longer than the cutoff, which can be estimated from the reference distribution (figure II.15, dark green vertical bar). As an example, if 50 % of the events in the reference distribution are shorter than the cutoff, then at the pause site there should likewise be an equal number of pause-free events below and above the cutoff. Therefore, the total number of non-paused events at the pause site is thus computed as twice the number of events shorter than the cutoff. The pausing efficiency is calculated as the remaining fraction of crossings, which must arise from the paused state.

In fact, this value is likely to be a lower bound on the true pausing efficiency since, contrary to our assumption, even events shorter than the cutoff may arise from the paused state with non-zero probability. Based on simulations, we estimate that true pausing efficiencies may be up to 15 % higher than the values we report, for the following reasons.

First, as shown in figure II.7, up to 6 % of the pauses occurring at every site will not occur in the ± 3 bp window around the expected position, and therefore the event most likely to be scored is a pause free event. We can estimate the *maximum* error caused by this effect by making the following assumptions: first, that 6 % of events in the crossing time distribution have no pause site in the window, and that all of these events are recorded as pause free.

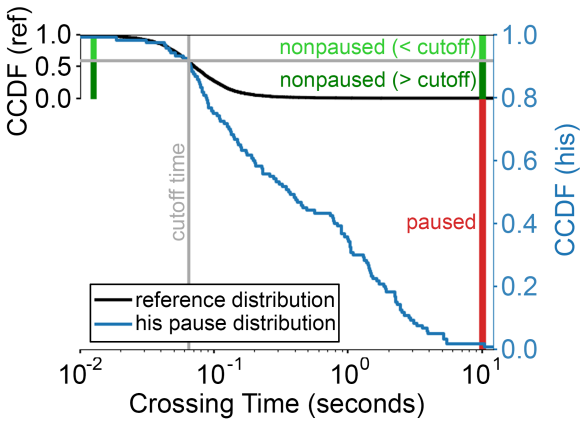


Figure II.15:
Description of the method used to calculate pausing efficiencies, illustrated for the 'his' pause at 10 pN assisting force. The reference distribution is rescaled to indicate the overlap between the two distributions below the cutoff time. For the distribution at the 'his' site, all events shorter than the cutoff are classified as nonpaused (light green), as well as events longer than the cutoff in the same proportion as in the reference distribution (dark green). The remaining events (red) are classified as paused.

Second, that the missed pause site crossings have the same pausing efficiency as the captured crossings.

Under those assumptions, the true pausing efficiency E is

$$E = \frac{N_{\mathrm{p,\,in}} + N_{\mathrm{p,\,out}}}{N_{\mathrm{tot}}} \tag{II.12}$$

where N_{tot} is the total number of crossings, $N_{\mathrm{p,in}}$ the number of paused crossings that occurred within the window, and $N_{\mathrm{p,out}}$ the number of paused crossings that occurred outside of the window—the latter being equal to $N_{\mathrm{tot}} \times 0.06E$. Meanwhile, the observed efficiency E_{obs} is only derived from paused crossings in the window, i.e.,

$$E_{\mathrm{obs}} = \frac{N_{p,in}}{N_{\mathrm{tot}}} = (1 - 0.06)E = 0.94E\,. \tag{II.13}$$

Therefore, an observed efficiency of 80 % will correspond to an efficiency of $\sim$ 85 %. However, it should be noted that the fraction of pauses occurring outside the window may be lower than 6 % as explained above, and there this effect may be smaller.

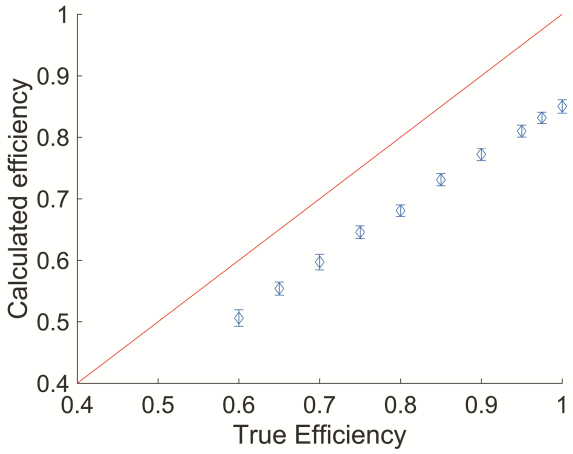


Figure II.16: Calculated efficiency vs. real efficiency in a simulated dataset

Second, our observed cutoff times are low, and at these time scales even datasets derived from completely different distributions (equivalent to 100 % pausing efficiency) may not always be distinguishable, particularly with the size of our datasets (150 to 200 events for pause sites, $\sim$ 2000 events for non-pause sites). To illustrate this, we generated synthetic datasets—one containing 2000 data points derived from the distribution of a maximum of six exponentials with an identical rate set to 30 bp/s (mimicking the pause-free sites crossing) and one containing 200 data points derived from a similar distribution, but with one of the exponents having a slower rate for a fraction of the data points (thus simulating the true efficiency). We selected a slow rate of 2 bp/s. We performed the same truncation algorithm as for the experimental data for different efficiencies. We found that in such a case, the calculated pausing efficiencies from our method underestimate the real efficiency by 10 % to 15 % (figure II.16). Similar results were obtained when using a gamma distribution with a shape factor of 2 and a scale factor of 0.25 to describe the pause crossing times (data not shown).

Finally, note that our pausing efficiencies should, in fact, be interpreted as "additional pausing efficiencies above the background amount of random pausing present in the reference data". However, random pausing events occurring outside the pause sites ("ubiquitous pauses") were rare compared to previous reports [108, 116]: the fraction of crossings in non-pause sites longer than 1 s (and therefore detectable as pauses by earlier methods) was < 0.2 % for assisting forces, reaching 1.8 % at 10 pN opposing force (figure II.17). We speculate that many pauses assigned as ubiquitous in previous studies with lower spatial accuracy were in fact sequence-specific but their exact position in the sequence could not be resolved. In comparison, the fraction of > 1 s crossings at pause sites was typically in the range of 5 %

to 35 % at assisting forces and 30 % to 50 % at opposing forces. Therefore, the assumption that no pausing occurs outside the pause sites would at most cause an additional slight underestimation of the true pausing efficiencies at the pause sites.

As a control we measured the efficiency of individual reference sites against all the other reference sites. As expected, reference sites showed pausing efficiencies around 0 (including "negative" efficiencies, corresponding to the case where events at a specific reference sites tend to be *faster* than the global reference distribution), and much lower than both major and minor pause sites (figure II.18).

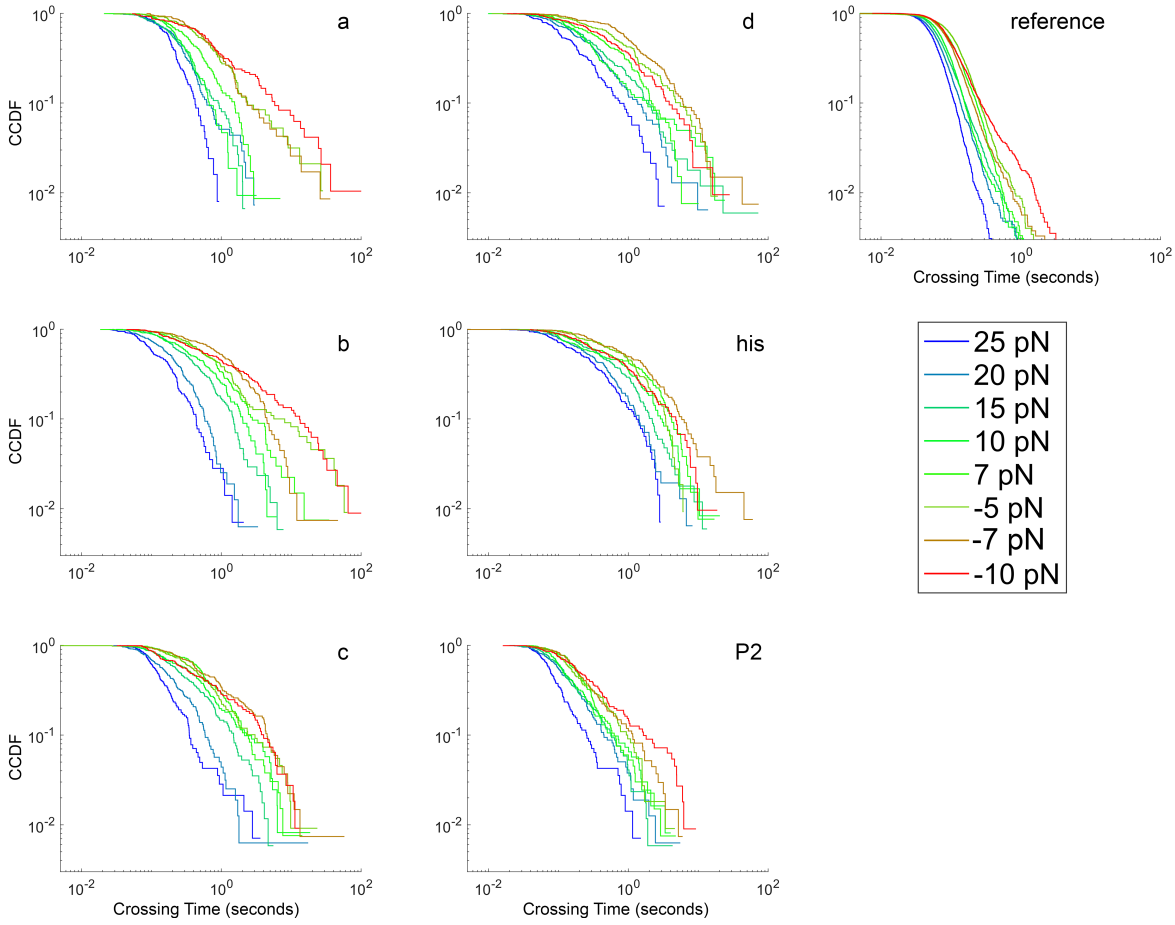


Figure II.17:
Crossing time distributions at various pause sites and at reference sites at different applied forces.

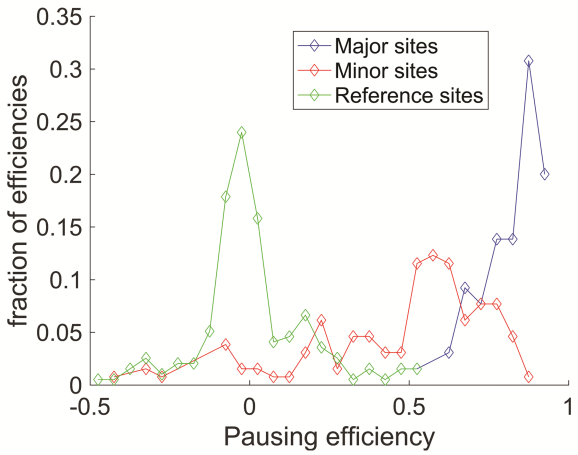


Figure II.18:
Calculated pausing efficiencies at different sites, calculated across the whole dataset.

Chapter II.3

Results

II.3.1 Nearly all RNAP molecules exhibit slow forward transcription rates at sequence-dependent pause sites

We computed the pausing efficiency for the sites 'a', 'b', 'c', 'd' and 'his' at different forces using the non-parametric method described above (figure II.19). All pause sites exhibit uniformly high pausing efficiencies ($> 70\,\%$ to $85\,\%$) that are independent of force. In other words, almost all RNAP molecules exhibit slower dynamics when crossing sequence-dependent pause sites, regardless of whether they entered an extended paused state. In contrast, the extrapolation-based method consistently underestimates the efficiency, particularly at high forces at sites 'a', 'b' and 'c'. Unlike pausing efficiencies, pause durations are strongly force dependent, as seen both in the distributions of pause site crossing times (figure II.17) and in the residence times (figure II.20).

II.3.2 Pause stabilization by backtracking occurs in two steps with distinct kinetics

Using the improved spatiotemporal resolution and positional accuracy of our method, we probed the dynamics of backtracking events down to 2 bp depths under opposing forces which are known to favor backtracking. Specifically, as illustrated in figure II.22a, we measured how far RNAP backtracked (backtrack depth), for how long the polymerase paused before it began to backtrack (pre-backtrack time), and how long it spent in the backtracked state (backtrack duration). Backtracking is highly site-specific, with the vast majority of backtracking events occurring at site 'b', and less frequently at site 'a' (figure II.22b).

We tested whether the sequence-dependence of backtracking could be explained by the changes in the thermodynamic stability of the transcription bubble as RNAP backtracks. For

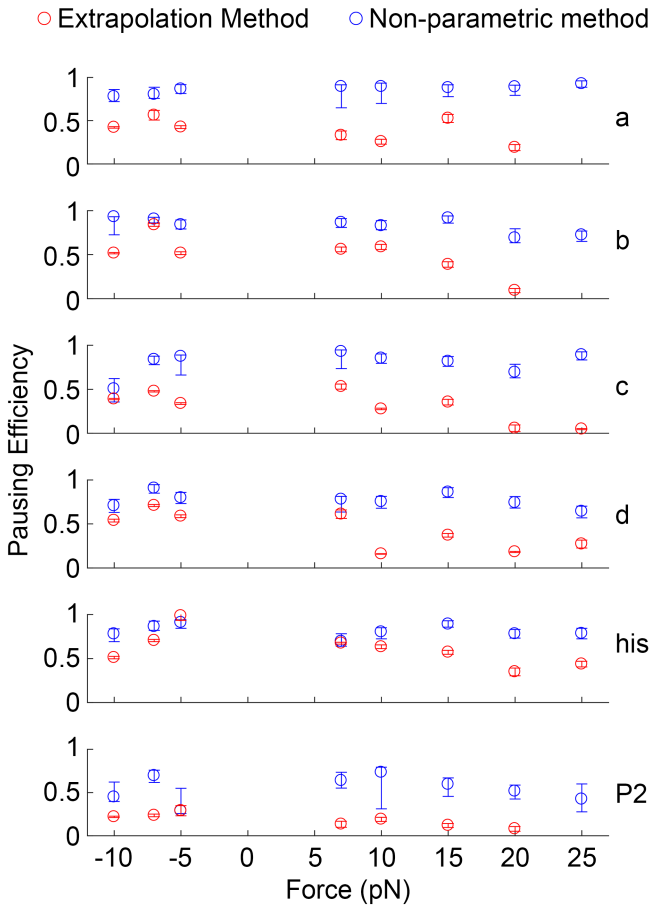


Figure II.19:
Pausing efficiencies at the major pause sites at different forces, calculated using extrapolation of the crossing time distributions above 1 s towards faster times (red), and by our nonparametric method (blue).

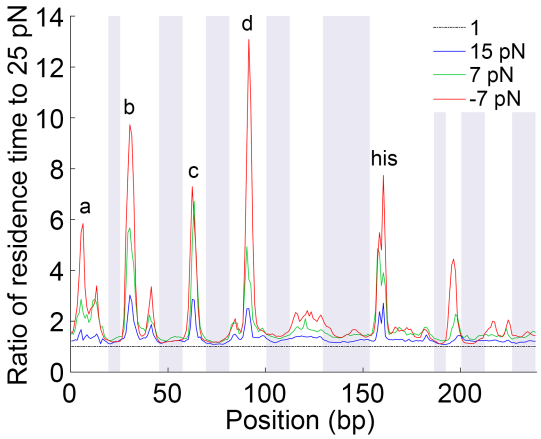


Figure II.20:
Residence time histogram ratios. The ratios of the residence times at three forces to the residence times at 25 pN assisting force are plotted. The ratio for 25 pN assisting force, which equals 1 by definition, is plotted as a reference.

every position along the template, we calculated the free energy change of forming of the transcription bubble, involving the melting of 12 bp of DNA hybrid (from the −12 position to the active site) and the formation of a 10 bp RNA-DNA hybrid, using tabulated nearest-neighbor free energy values [117, 118]. We then calculated the free energy change associated with backtracking by up to 8 bp at each of the pause sites (figure II.21a, left). Only sites 'a', 'b' and 'his' exhibited an energetic gain from backtracking without the need to cross a significant energetic barrier. This observation suggests that a significant part of the large backtracking propensity at sites 'a' and 'b' may be attributed to the gain in free energy resulting from moving the transcription bubble backwards. Although backtracking at the 'his' site appears energetically favorable in this simplified model, it is most likely inhibited by the hairpin in the nascent RNA. More generally, additional sequence-dependent factors that may affect the backtracking propensity are other secondary structures in the nascent RNA [109] and interactions between the nucleic acids and the polymerase itself.

To test whether bubble thermodynamics contribute to backtracking propensity in the bacterial cell, we analyzed the pause sites mapped by Imashimizu et al. [12]. We used the same technique to calculate the bubble free energies at the pause sites (figure II.17a, right). At the sites mapped for WT *E. coli*, each of the first three steps were energetically favorable at the majority (at least 55 %) of the sites ($p < 0.02$, two-sided binomial test) with the least favorable step (the first one) averaging 0.16 kcal/mol across all sites. At the sites mapped for Δ*greAB E. coli*, each of the first seven steps were energetically favorable at the majority (at least 58 %) of the sites ($p < 0.01$, two-sided binomial test) with the least favorable step (the seventh)

averaging 0.14 kcal/mol across all sites. Conversely, we found that across all annotated transcribed sequences in the *E. coli* genome (`NC_007779.1`), sites where backtracking is energetically favorable are in a slight minority (49.95 %, $p < 0.05$, two-sided binomial test). Overall, these findings support a model in which the differences in the energetic cost of opening a bubble at various sites are a major contributor to the difference in backtracking propensity. This result is consistent with recent findings from genome-wide analysis of pausing in bacteria and yeast [119].

We further characterized the backtracking dynamics at site 'b'. First, the backtrack depth and duration were positively correlated (figure II.22c) with a sub-linear dependence that points to a diffusive backtracking mechanism [48, 120]. Return of the polymerase to the active site does not necessarily imply recovery from the paused state—in many backtracking events, the polymerase successfully moves back towards the active site, only to backtrack again once or several times before actual recovery (figure II.22a). Second, we found that RNAP does not begin to backtrack immediately upon entering a pause. Instead, it takes at least a second before the enzyme begins to move backwards (with most backtracking events starting 1 s to 10 s after the beginning of the pause, figure II.22d). This observation indicates that the stabilization of a pause by backtracking occurs in two steps: 1) rapid formation of an initial paused state, which is either non-backtracked or backtracked by 1 bp at most, and 2) slow conversion into a deep and long-lived backtracked state.

To further characterize the backtracking process, we also conducted experiments in the presence of elongation factor GreB (at a concentration of 0.87 μM), which rescues elongation complexes backtracked by as little as 2 base pairs, but inhibits transcription by non-backtracked RNAP [120] (figure II.22 and figure II.21). Transcription data collected in the presence of GreB at 10 pN opposing force, at which backtracking is most favored, displays shorter and less deep backtracking events, with rapid recovery indicative of transcript cleavage and transcription directly from the backtracked position, in contrast to the slower, diffusive return observed in the absence of GreB [88]. GreB has a slight opposite effect on non-backtracked pauses: GreB slightly increased the crossing times at all time scales at the sites 'c', 'd', and 'his', consistent with the low degree of backtracking observed at those sites (figure II.23). The effect of GreB was different for sites 'a' and 'b' (figure II.23)—at short time scales, crossing times at 'a' and 'b' were unaffected, or even slightly increased in presence of GreB; however, pausing events longer than $\sim$ 3 s, comprising 20 % to 25 % of the events, were highly attenuated. Pause-free sites display similar behavior to pauses 'a' and 'b', but only $\sim$ 3 % of the crossings (corresponding to time scales > 0.7 s) are shortened by GreB, indicating that backtracking outside the main pause sites occur at a very low frequency. This result further confirms that ≥ 2 bp-backtracked states are formed slowly from non-backtracked or 1 bp backtracked paused states.

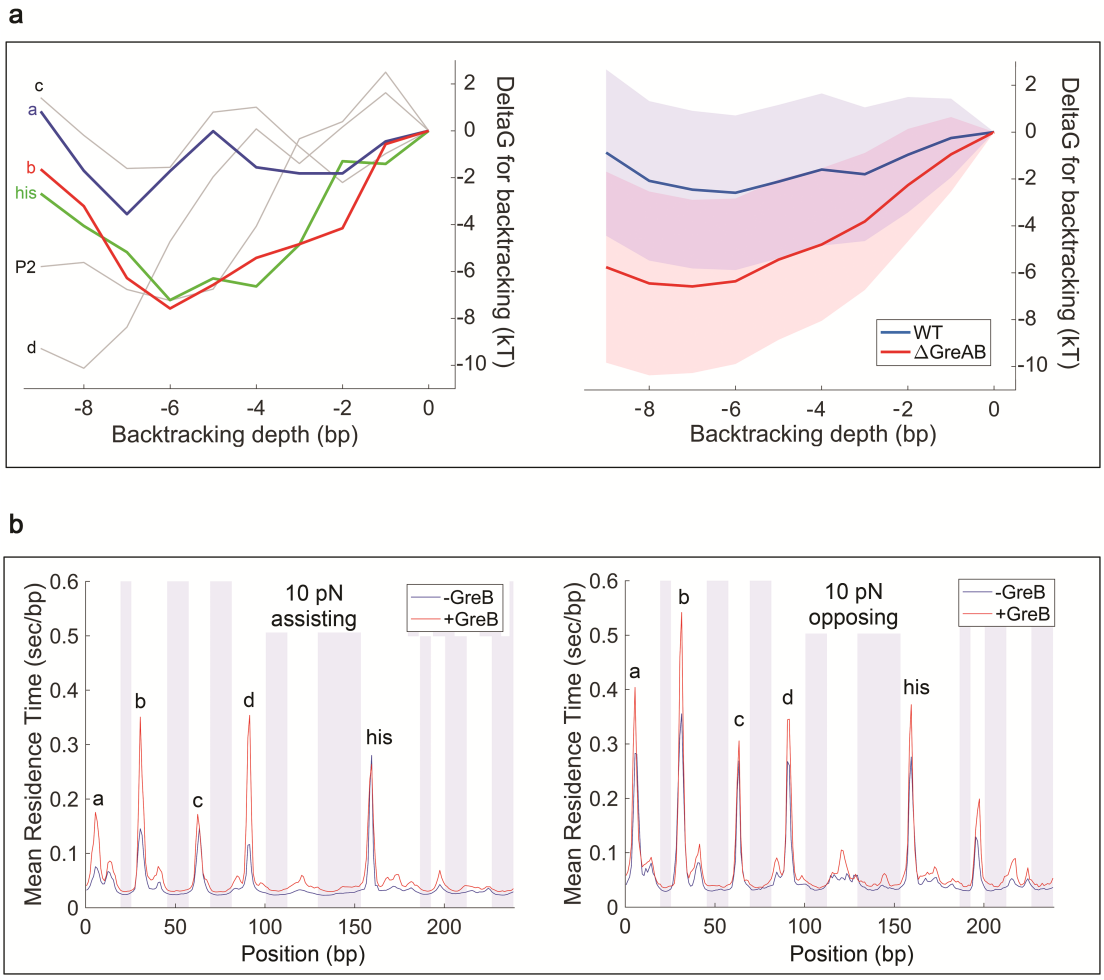


Figure II.21:

(a) (Left) Calculated change in free energy of the transcription bubble during backtracking at different pause sites. Energies were calculated for a 10 bp RNA:DNA hybrid and a 12 bp transcription bubble. At sites 'a' and 'b', backtracking is energetically favorable to at least 3 bp to 4 bp without encountering significant barriers, but not at sites 'c', 'd', and 'P2'. The 'his' site also shows a large energetic gain, which is most likely offset by the hairpin in the nascent RNA. (Right) Identical calculation averaged across the backtrack-prone sites mapped by Imashimizu et al., for both the WT (blue) and the Δ*greAB* (red) datasets. Solid line, mean difference; shaded area, one standard deviation.

(b) Effect of GreB on transcription dynamics at 10 pN assisting force and 10 pN opposing force.

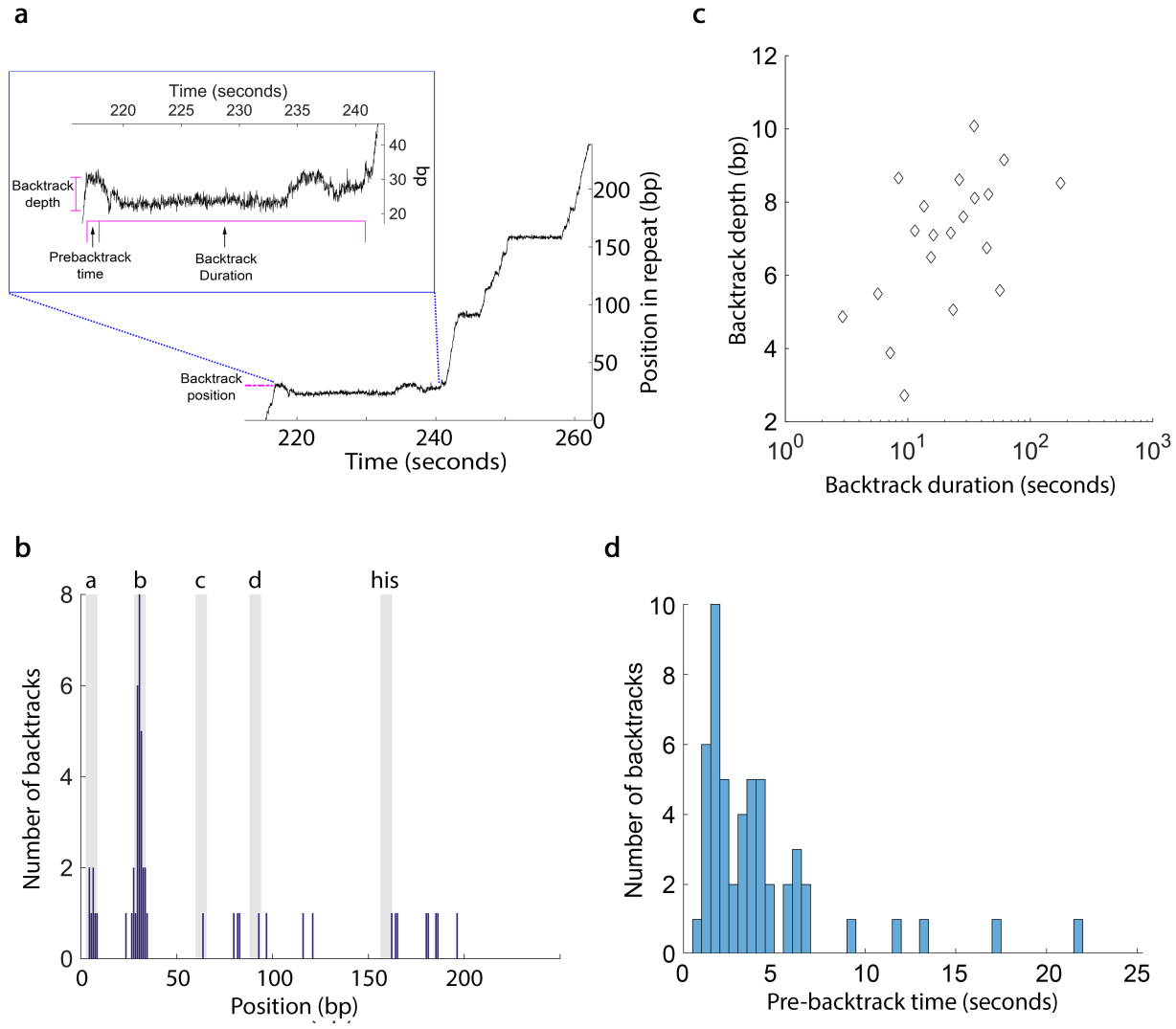


Figure II.22:

Analysis of backtracking events.

(a) The trace shown contains a backtracking event occurring at site 'b'. The backtrack depth, pre-backtrack time and backtrack duration are labeled.

(b) Histogram of backtracking events by position.

(c) Backtracking depths and times measured at opposing forces for site 'b'.

(d) Histogram of pre-backtrack times measured at site 'b' at opposing forces.

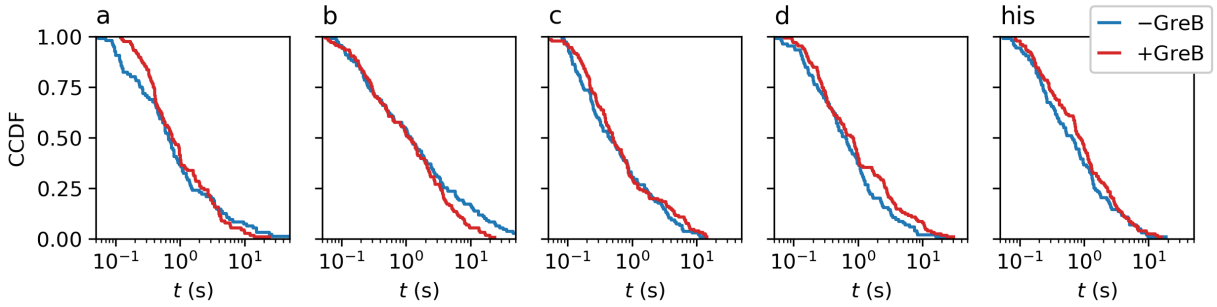


Figure II.23:
Effect of 0.87 µM GreB on the crossing time distributions at the major pause sites, measured at 10 pN opposing force. At short time scales, comprising 80 % to 90 % of the measured events, GreB slightly increases crossing times. Therefore, mean residence times are longer in the presence of GreB at pause sites (figure II.21). At sites 'a' and 'b', for the longest events (20 % to 25 %), GreB reduces the crossing times, indicating these are backtracked events.

II.3.3 Nascent-RNA hairpins enhance or attenuate pausing depending on the sequence context and the applied force

We probed the effect of nascent-RNA on the transcriptional dynamics by addition of 0.1 mg/ml RNase A [109]. Consistent with reports that pausing is stimulated by the nascent RNA hairpin at the 'his' site, RNase strongly attenuated, but did not abolish the pausing at that site (figure II.25a). We found that nascent RNA also affects pausing dynamics in other sites, and that the direction and magnitude of the effect are sequence dependent. Pause 'd' was attenuated, though to a smaller extent than 'his', whereas the otherwise weak pause 'P2' was strongly enhanced by RNase. Modulation of pausing by the nascent RNA and backtracking appear to be mutually exclusive, as the backtracking-prone sites 'a' and 'b' did not exhibit sensitivity to RNase. The residence times at several of the minor pause sites (such as 'P1' and 'P6') were also modulated by removal of the transcript.

Next, we analyzed how the applied force changed the RNase sensitivity of the affected pause sites (figure II.25b, c). The effect of RNase was consistently stronger at opposing force than at assisting force. This observation could be explained by two scenarios. First, nascent RNA structures may interact more strongly with RNAP in the pre-translocated state, which is favored by opposing force. Second, under opposing force, transcription rates are lower, which may give more time for RNA structures to form, thus enhancing their effect. For the

pause sites 'a', 'b', 'c', 'd', 'his', and P2, we simulated the cotranscriptional folding of 50 RNA bases with the 3' end at the −11 position relative to the pause site (immediately upstream of the exit channel) using Kinefold [121] (figure II.24). We performed the simulation using two transcription rates: 29.1 ms/base (rate at 10 pN assisting force) and 35.3 ms/base (rate at 7 pN opposing force). For every simulation, we checked whether among the five most stable structures there is a structure containing a helix that ends immediately upstream of the polymerase at the pause site (positions −15 to −11). We plotted the helix trace for such helices (percentage of formation of the helix as a function of time). The results are shown in figure II.21. For the sites 'P2' and 'his' a stable helix is formed at least 40 ms before RNAP reaches the pause site, with for site 'd' a less stable structure appears ∼ 80 ms before RNAP reaches the pause site. The applied force, through its effect on the transcription rate, does not cause any significant changes to the stability of the formed structural or to the timing of their formation relative to the arrival of RNAP at the pause site. Overall, these simulations indicate that RNA folding is likely to be fast compared to transcription and therefore the effect of RNA on pausing is unlikely to change due to the small (∼ 15 %) variation in transcription rate over the range of forces tested. Accordingly, we tend to favor the hypothesis that nascent-RNA structure interacts predominantly with RNAP in pre-translocated state.

Previous studies using bulk transcription assays have found that the mutating the nascent RNA hairpin at the 'his' site reduced pause durations with minimal effects on measured pausing efficiencies, while mutations to the consensus pause elements reduced both [49, 122–124]. However, it is unclear whether this is due to limited temporal resolution (∼ 10 s) that precluded the detection of short pauses. Using the enhanced resolution in our assay, we tested the effects of RNase on pausing efficiencies at the 'his', 'd', and 'P2' sites. We found that in contrast to its effect on the pause residence times, the effect of RNase on pausing efficiency is minimal, with no significant changes at assisting force and only a ∼ 25 % reduction at opposing force for the 'his' site. These observations confirm that the interaction between the nascent RNA and RNAP plays only a minor role in pause entry, and primarily serves to enhance (for 'his' and 'd') or inhibit (for 'P2') the formation of longer lived paused states.

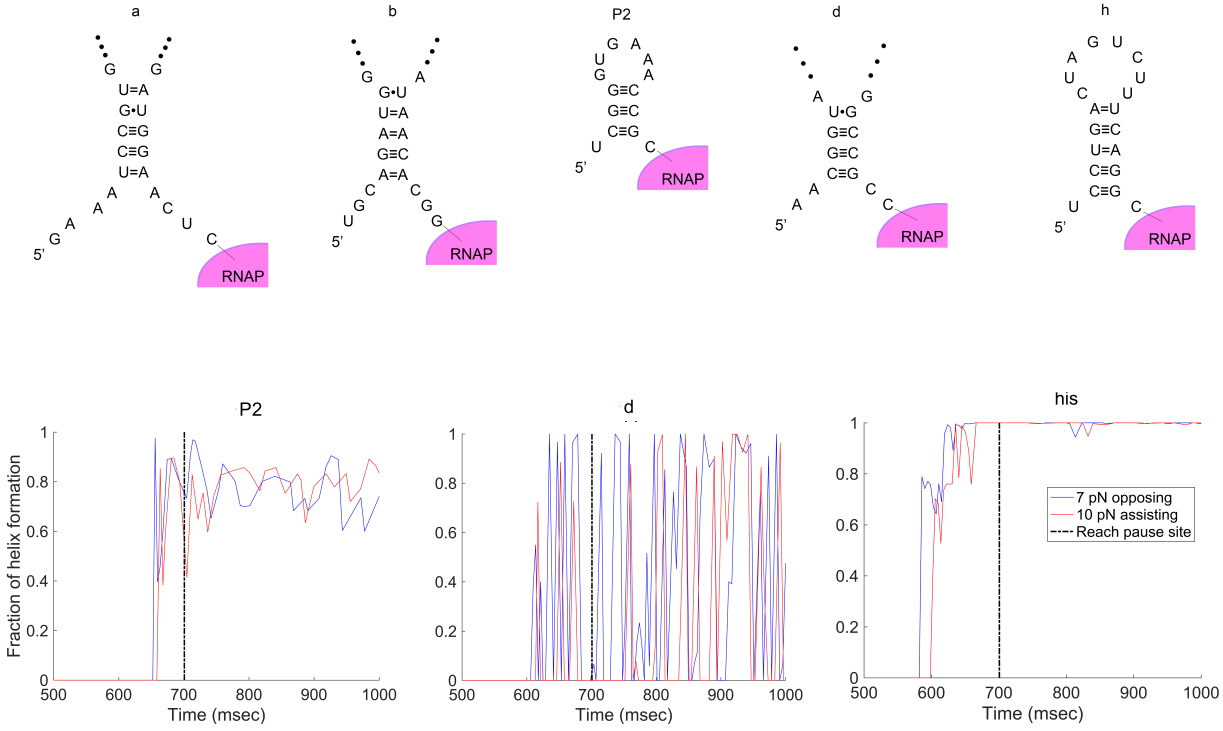


Figure II.24:

Kinefold analysis of RNA structures in the pause sites. Top: Representative RNA structures formed near the exit channel when RNAP is located at the pause site. Bottom: Helix traces for the structure shown for sites P2, d and 'his'. The black dashed line indicates the completion of transcription of the 50 base RNA, equivalent to the arrival of RNAP at the pause site, thereby releasing the −11 RNA base from the exit channel.

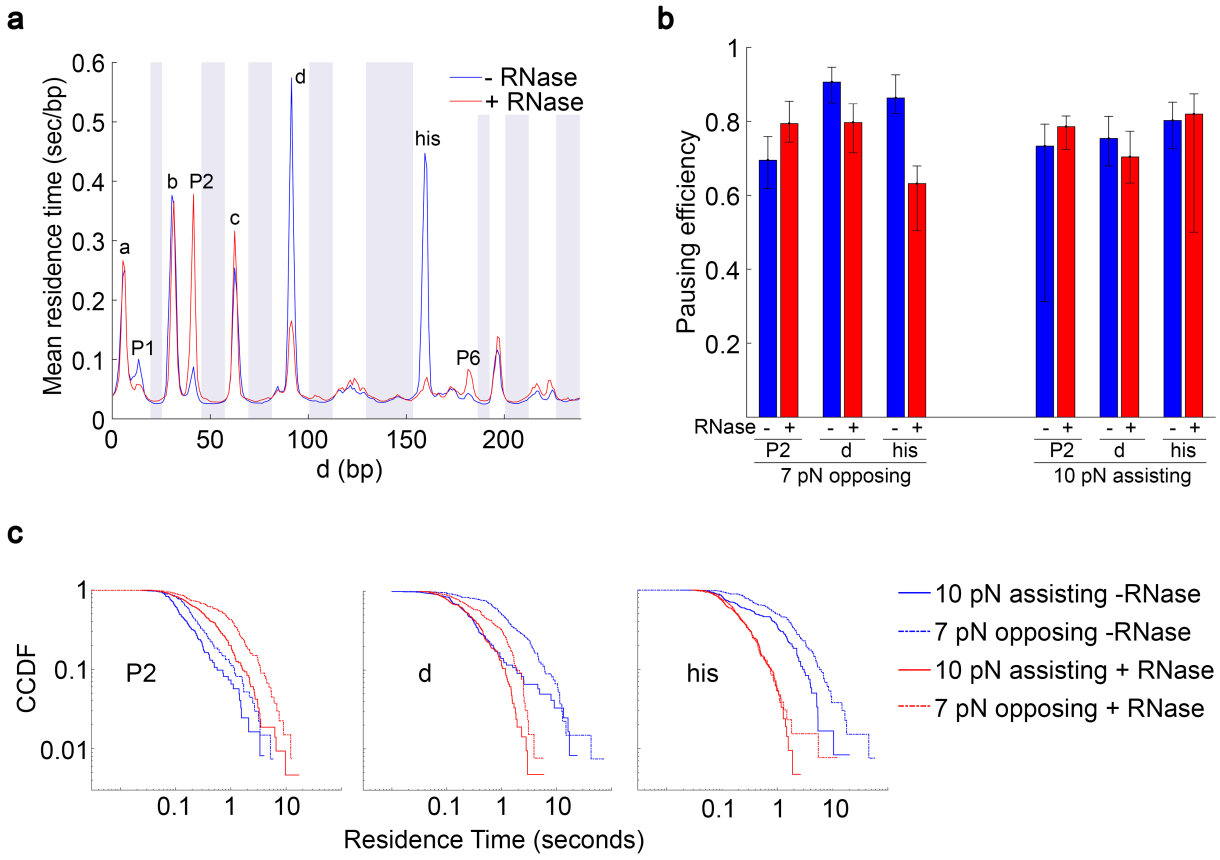


Figure II.25:
Effect of RNase on pausing dynamics.

(a) Residence time histograms collected at 10 pN assisting force with and without RNase.

(b) Effect of RNase on pausing efficiencies.

(c) Effect of RNase on residence time distributions at sites 'P2', 'd' and 'his'.

Chapter II.4

Discussion

Sequence-specific pausing involves the formation of elemental paused states that are stabilized by processes such as backtracking and RNA hairpin formation. The enhanced temporal resolution of our assay revealed that in addition to these mechanisms, pause sequences facilitate pause entry by reducing the forward transcription rate of RNAP: nearly all RNAP molecules exhibit slow transcription dynamics when crossing a pause site, even under conditions in which few or no pausing events are long enough to be detected directly using previous methods. This most likely takes place through sequence-dependent inhibition of forward translocation, as evident from the strong force-dependence of pause durations. This is further supported by the inhibition of forward translocation of *E. coli* RNAP by consensus pause elements [11] and by studies of *S. cerevisiae* RNA polymerase II, for which sequence-specific translocation barriers that inhibit forward motion of the polymerase have been implicated in pausing [125]. We propose that inhibition of the on-pathway elongation dynamics of RNAP allows time for transitions into stable off-pathway paused states, such as hairpin-stabilized or backtracked-stabilized states [126] (figure II.26).

The universally high and force-independent pausing efficiencies can also be rationalized by the existence of a paused state which is accessed at a very high rate relative to forward transcription, resulting in high pausing efficiencies at all forces. From a functional perspective this model is equivalent to an effective reduction of the on-pathway rate.

Previous studies of transcriptional pausing were blind to the dynamics in the 100 ms to 1000 ms time scale [48, 108] and estimated the number of pauses shorter than 1 s by extrapolating from the distribution of events longer than 1 s, usually assumed to be exponential. This approach implicitly assumes that the rate of pause escape is the same in both time scales; in fact, it yields a highly inaccurate picture of the pausing dynamics at short time scales (figure II.13). The "slowest-crossing" method presented here for computing crossing times enhances the temporal resolution of the dynamics at pause sites, down to ~ 100 ms. The non-parametric method for computing the pausing efficiencies presented here relies on the more conservative assumption that pause-free crossings at the pause sites occur via the

same mechanism as crossings of non-pause sites and therefore follow the same distribution. Using this approach, we found that pausing efficiencies are much higher than previously determined and are independent of force.

The enhanced temporal resolution of our assay enabled a detailed characterization of the roles played by backtracking and nascent RNA structure in pausing dynamics. Structural studies have found differences between the conformations of 1 bp backtracked and deeper backtracked elongation complexes, both for bacterial and eukaryotic RNAP [127, 128]. However, it was unclear whether these structural differences manifest in a functional effect on the dynamics of backtracking. Our results provide direct evidence that the formation of $\geq$ 2 bp backtracks involves the rapid and efficient formation of an initial paused state (non-backtracked or 1 bp backtracked) followed by a slower formation of deep backtracked states (figure II.26). Optical trapping studies on nucleosome-induced pausing have also yielded indirect evidence of similar behavior by yeast RNA polymerase II [129]. While it is unclear whether the initial paused state is a non-backtracked or a 1 bp backtracked state, we propose that the two-step nature of backtracking may be a general property of elongation complexes.

As with backtracking, the effects of RNA structure on pausing are site-specific; they vary both in their direction and in their magnitude (figure II.26). While pausing at the 'his' site is strongly dependent on the hairpin, site 'd' retained significant pausing in the presence of RNase, and the RNA structure primarily stabilizes paused states longer than 1 s. At site 'P2', pausing is inhibited by the nascent-RNA. Studies of eukaryotic RNA polymerase II [109] have suggested that RNA structures diminish pausing by generating a physical barrier for RNAP backtracking [130]. However, we observed no significant backtracking at site 'P2', and RNase had no effect on backtracking in general or on the backtrack-prone sites 'a' and 'b'. The CTD at 'P2' was also not affected by GreB, and the effect of RNase was also observed at assisting force, where backtracking is not favored. We therefore conclude that the nascent RNA inhibits pausing at site 'P2' by a distinct interaction with RNAP, and not by inhibiting backtracking.

Our results contrast with previous reports that detected no effect of nascent RNA folding on pausing [50], and highlight the importance of sequence resolved, high-resolution studies of pausing. The low temporal resolution and inability to resolve the position of RNAP in previous methods would cause the diverse effect of nascent RNA to average out. Sequence resolved characterization of pausing at high temporal resolution revealed a far more complex picture of transcriptional regulation by nascent RNA structure. Since the nascent transcript can bind species such as ribosomes [131] and termination factors [132], it may serve as a fine-tuning element in the transcription cycle, enabling flexible modulation of elongation rates in a context-dependent manner.

Transcriptional pauses play a crucial role in the regulation of gene expression and in the coordination of transcription with other processes. Understanding the molecular transitions

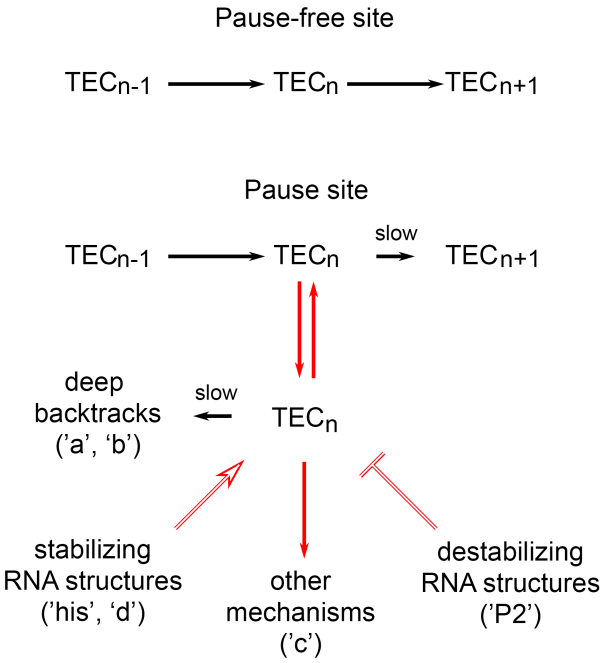


Figure II.26:
Proposed model for transcriptional pausing by *E. coli* RNAP. TEC: Transcription Elongation Complex; the indices $n-1$, n and $n+1$ indicate the length of the RNA product. At pause sites, the paused state is rendered kinetically accessible to the polymerase through a slowing down of the on-pathway forward translocation rate. Depending on the sequence context, this paused state can transition slowly to a ≥ 2 bp backtracked state (in sites 'a' and 'b'), be stabilized or destabilized by the nascent RNA ('his', 'd' and 'P2') or be stabilized by other mechanisms (such as in site 'c', which exhibited neither backtracking nor RNase sensitivity).

that lead from pause-free transcription to paused states requires tools that permit the characterization of pausing dynamics at high spatiotemporal resolution. The development of such tools in this work resulted in valuable insights into the mechanism of pausing and opens the door to more detailed studies on pause entry of both bacterial and eukaryotic RNA polymerases.

Chapter II.5

Materials and methods

All DNA modifying enzymes were purchased from New England Biolabs. Oligonucleotides were purchased from IDT. Nucleotide triphosphates were purchased from Thermo Scientific, and standard salts and buffer components were purchased from Sigma Aldrich. Carboxylated 1 µm polystyrene beads were purchased from Bangs Labs.

II.5.1 Plasmids and DNA templates

Plasmids pIA1127 (for expression of σ^{70}), pIA1234 (for expression of sortagged RNA polymerase), and pIA2-6 (used as a template for preparing DNA handles) were a gift from Irina Artsimovitch. Plasmid for the expression of sortase was a gift from David Liu.

The template was derived from a plasmid containing the T7A1 promoter, a $\sim$ 1 kb downstream spacer region, eight repeats containing the 'his' pause, and finally an *rrnB* T1 terminator sequence [9].

II.5.2 Preparation of DNA template, enzymes, beads, and stalled complexes

The DNA template was prepared as described in part I, with the single modification that opposing force templates were obtained by treatment with Klenow 3'-5' exo− polymerase (1 unit per µg DNA) and 0.1 mM ddCTP (instead of the ddATP used for assisting force template).

σ^{70} preparation, RNA polymerase holoenzyme preparation, sortagging, and biotinylation, bead coupling to oligos, bead passivation, and stalled complex preparation were performed as described in part I.

II.5.3 Preparation of GreB

The gene for GreB was cloned into a pET vector by ligation independent cloning (Addgene #29653). The plasmid was transformed in Rosetta2 cells, the bacteria were grown in 1 liter of 2YT medium supplemented with 1 % glucose, NPS (25 mM $(NH_4)_2SO_4$, 50 mM KH_2PO_4, 50 mM Na_2HPO_4), 1 mM magnesium sulfate, 34 µg/ml chloramphenicol and 50 µg/ml kanamycin. The culture was grown at 37 °C to an OD_{600} of 0.6, IPTG was added to 0.5 mM and transformation proceeded for 4 hours at 37 °C. The bacteria were then centrifuged, and dispersed in 40 ml of lysis buffer (Tris 100 mM pH = 7.9, 25 mM imidazole, 1 M NaCl, 2 mM β-mercaptoethanol) supplemented with 1 mM PMSF and 0.2 mg/ml lysozyme. The bacteria were lysed by sonication, and the solution was centrifuged and filtered.

The sample was loaded on a 2 ml Ni-NTA column, washed with 12 ml of lysis buffer, followed by 12 ml of lysis buffer with 50 mM imidazole, and finally eluted with lysis buffer with 300 mM imidazole. TEV protease was added at a 1:10 molar ratio, and the sample was incubated overnight at 4 °C while dialysing against lysis buffer. The sample was passed again over 1 ml Ni-NTA beads, concentrated to < 3 ml, and loaded on a sephacryl S100 gel filtration column equilibrated with Tris 25 mM pH = 8, 1 M NaCl, 1 mM EDTA, 1 mM DTT. Fractions containing clean GreB were pooled and concentrated to ~ 50 µM; glycerol was added to 50 %; and the protein was flash frozen with liquid nitrogen and stored at −80 °C.

When performing experiments with GreB, the protein was dialyzed first into HEPES 25 mM pH = 8, 1 M KCl, 1 mM DTT and 1 mM EDTA so that it could be mixed into the experimental buffer in precomputed ratios in order to maintain the buffer composition.

All proteins were > 95 % pure based on SDS-PAGE. Holo-RNAP activity and pausing was confirmed using short template containing a T7A1 promoter, 29 bp U-less cassette and a downstream 'his' site. GreB activity was tested by the rescue of a 2 bp backtracked elongation complex assembled using an RNA oligonucleotide with two mismatched bases at the 3' end [133].

Part III

Unraveling the thousand word picture: an introduction to super-resolution data analysis

This review of super-resolution analysis techniques was originally written as a part of

> Unraveling the thousand word picture: an introduction to super-resolution data analysis.
>
> A. Lee, K. Tsekouras, C. Calderon, C. Bustamante, S. Pressé. *Chem. Rev.* (2017), *117*, 7276–7330.

It has been reproduced here, with modifications, with the permission of all authors.

For this work, Carlos Bustamante and Antony Lee acknowledge support from the Nanomachines program (KC1203) funded by the Director, Office of Science, Office of Basic Energy Sciences of the U.S. Department of Energy (DOE) contract no. DE-AC02-05CH11231 (step-finding algorithms), by the National Institute of Health grants R01GM071552 and R01GM032543 (fluorescent protein characterization), and by the Howard Hughes Medical Institute (fluorescent protein counting). Steve Pressé acknowledges the support of NSF MCB 1412259 as well as startup from IUPUI and ASU. Christopher Calderon was supported by internal R&D funds from Ursa Analytics, Inc. The authors thank Ioannis Sgouralis for many helpful suggestions.

Processes fundamental to life, including DNA transcription, RNA translation, protein folding, and assembly of proteins into larger complexes, occur at length scales smaller than the diffraction limit of light used to probe them (< 200 nm). For this reason, up until a decade ago, these processes were largely inaccessible to conventional microscopy methods. Key technical achievements by way of experiments, from structured illumination methods [134, 135] to manipulations of fluorophore photophysics [56–58], have peered into this previously impenetrable scale with several techniques now providing detailed in vivo 3D images.

On the experimental front, many technical challenges remain including the following: high density labeling; poor time resolution at the expense of high spatial resolution; challenges with fluorophore activation and complex photophysics; overexpression of select proteins altering cell homeostasis; and high light intensity, some $\sim 10^4$ times higher than that under which cells have evolved for methods such as photoactivated localization microscopy [136]. Despite these challenges, experiments have begun to resolve the spatiotemporal dynamics and organization of cellular components within their native environment, revealing, for instance, the intricacy of yeast DNA transfer from mother to daughter cell [137] and the stochastic assembly of chemoreceptors on *E. coli*'s surface [138]. What is more, recent advances in optics have mitigated the spatial-temporal resolution trade-off providing greater in vivo resolution in 3D [139–147]. Advances continue to accrue, with the latest techniques reaching spatial resolutions of ~ 1 nm and temporal resolutions on the order of microseconds [148].

Ten years have passed since the inception of super-resolution microscopy and the variety of data collected has presented new modeling challenges [149]. Initial data analysis methods, such as mean square displacement analyses, were directly motivated from the analysis of bulk ensemble data largely inspired by Occam's razor. Thus, such methods did not explicitly take advantage of the richness of single molecule datasets such as their temporal ordering or even their intrinsic heterogeneity.

A large fraction of this review is devoted to later "data-driven" efforts, deeply inspired from the fields of machine learning and inference, and increasingly available through an array of open-source software [150–154], to turn the thousand-word picture provided by super-resolution methods into a quantitative narrative.

Here, after presenting the basic physics of super-resolution methods (chapter III.1), we tackle two fundamental challenges for the analysis of data generated by such methods: the localization problem (chapter III.2) and the counting problem (chapter III.3).

Chapter III.1

Beating the diffraction limit: an introduction

III.1.1 Why fluorescence microscopy?

Upon excitation of a sample within a specific wavelength range (the absorption spectrum), a fluorophore emits light at a longer wavelength (the emission spectrum). The excitation wavelength may be filtered away leaving behind only the emission from the fluorescent components. In this way, fluorescence brings improved contrast to microscopy.

The first fluorescence microscopes, developed by the Carl Zeiss company and others in the early 20th century, relied either on the autofluorescence of various tissues or chemical dyes and stains such as fluorescein [155]. An important milestone in increasing the ability to fluorescently label a given biological structure was achieved by Coons et al. in the 1940s, who demonstrated that antibodies, raised to bind a specific antigen with high specificity, could be attached to fluorescent dyes, thus realizing a method to fluorescently label any antigen of interest [156]. The subsequent discovery of the green fluorescent protein [157], together with advances in molecular biology techniques, then allowed the expression of proteins directly fused to fluorescent markers by the end of the 20th century [158].

At the same time, the detection of the signal from single fluorophores (rather than larger labeled structures) was achieved by progressive improvements in instrumentation [159, 160]. This powerful combination of new advanced optical techniques with fluorescent protein tags, which could be detected in live cells at the single molecule level, set the stage for a new era of measurements in cell biology and biophysics [147].

III.1.2 Point spread functions and the diffraction limit

Although labeling techniques have greatly improved over the last century, fundamental physical reasons have limited the resolution achievable by optical microscopy. Historically, this resolution has been defined as the ability to distinguish two close objects.

As early as 1834, Airy derived the profile of the diffraction pattern, or point spread function (PSF), of a point source of radiation imaged through a telescope, now known as the Airy disk [161]. He established that "the image of a star will not be a point but a bright circle surrounded by a series of bright rings. The angular diameter of these will depend on nothing but the aperture of the telescope, and will be [*sic*] inversely as the aperture" [161]. More precisely, for a telescope of aperture a imaging at a wavelength λ, the intensity I at an angle θ from the optical axis, relative to the intensity I_0 at the center, is given by

$$I(x)/I_0 = \left(2J_1(x)/x\right)^2 \tag{III.1}$$

where $x = (2\pi/\lambda)\, a \sin\theta$ and J_1 is the first order Bessel function of the first kind. Rings appear at the maxima $x = x_1, x_2, ...$ of $I(x)$. In the limit of small angles (i.e., $\theta \approx \sin\theta$), these maxima correspond to $\theta_i = \lambda x_i/(2\pi a)$. Thus, the angular diameters of the rings are indeed inversely proportional to the aperture a figure III.1.

A few decades later, Abbe showed that a similar result held for optical microscopy: a point source imaged at a wavelength λ through a microscope objective of numerical aperture NA, defined as the product of the index of refraction of the medium between the objective and the sample, n, and the sine of the half angular aperture of the objective, θ, yields a spot of size $d \approx \lambda/2\mathrm{NA}$ in the transverse direction and $2\lambda/\mathrm{NA}^2$ in the axial direction [162] figure III.2.

Whether in astronomy or microscopy, it is the finite extent of the image of a point source that limits our ability to separate two objects nearby. In 1879, Rayleigh suggested a rule, now called the Rayleigh criterion, whereas two diffraction spots could be considered as resolved if their centers were further apart than the center of a spot is from its first zero in intensity [163] (figure III.3). He emphasized that this rule was simply suggested as an approximation "in view of the necessary uncertainty as to what exactly is meant by resolution", though this rule still remains in use today [164]. In fact, it is generally agreed in astronomy that spots up to $\sim 20\,\%$ closer are resolvable [164].

Nowadays, super-resolution imaging continues to leverage ideas and tools from astronomy, both on the experimental [165] and analysis side [166].

Even though the Rayleigh criterion may not be strictly accurate, the resolution of a microscope is certainly inversely correlated with the size of the diffraction spot. As this spot has a size of $d = \lambda/2\mathrm{NA}$ in the transverse direction and $d = 2\lambda/\mathrm{NA}^2$ in the axial direction, improvements to the resolution are achieved by working at a shorter wavelength or larger numerical aperture.

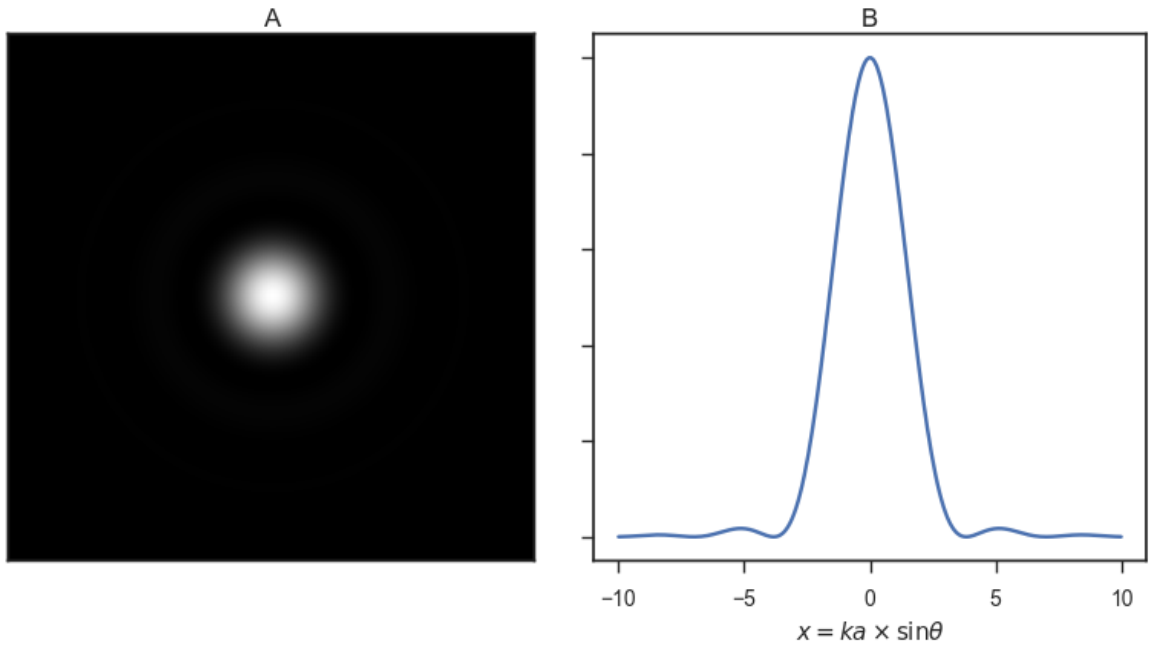


Figure III.1:
A point emitter generates an Airy spot (a) with an intensity profile (b) given by equation III.1. The wavenumber k, used in (b), is $2\pi/\lambda$. The intensity profile and diffraction spots were plotted using a simple Python script.

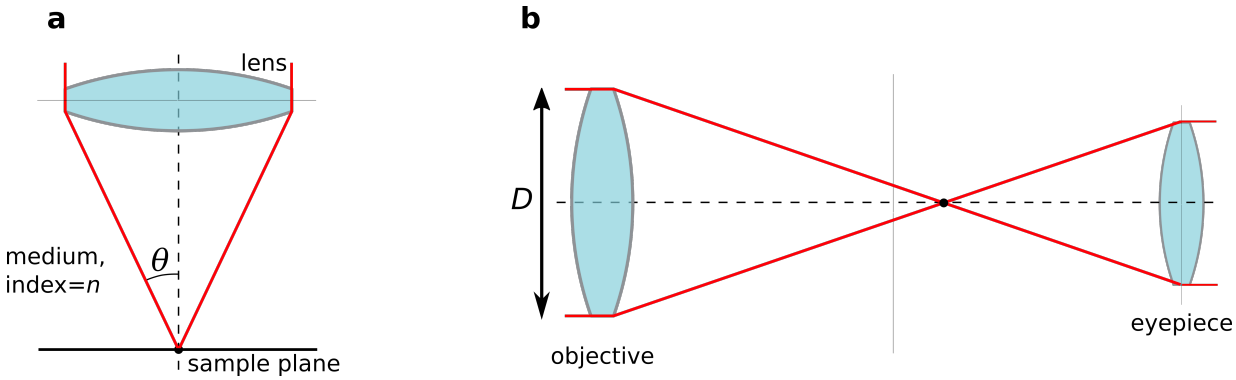


Figure III.2:
Microscope seen as a telescope. (a) A microscope's resolution is determined by the numerical aperture NA of its objective, which is defined as the product of the index of refraction of the medium between the objective and the sample, n, and the sine of the half angular aperture, θ. (b) A telescope's angular resolution is determined by its (physical) aperture, D.

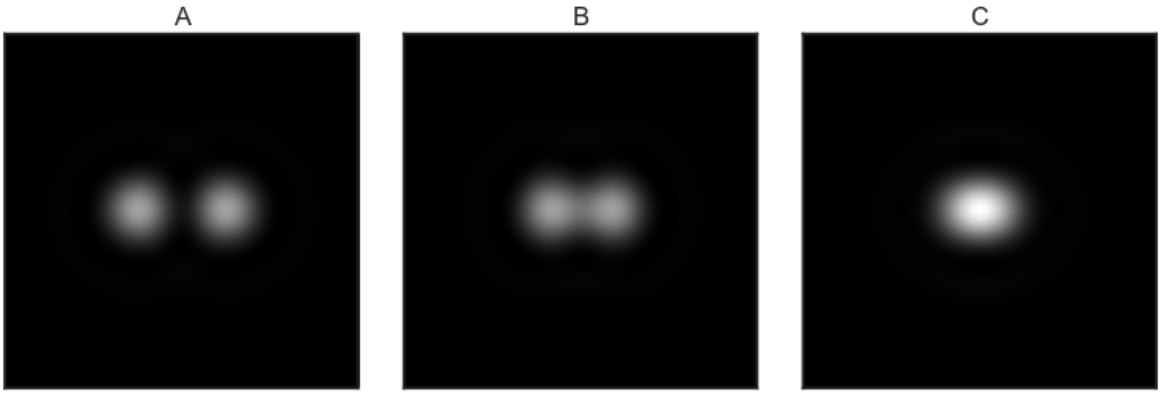


Figure III.3:
(a) Fully resolved, (b) barely resolved, and (c) non-resolved Airy diffraction spots according to the Rayleigh criterion. The intensity profile and diffraction spots were plotted using a simple Python script.

The room for improvement from changes in the wavelength is limited by the spectrum of visible light, $\lambda = 400$ to 700 nm. Electron microscopy achieves a much higher, near-atomic resolution by operating at a pm-scale wavelength, but this comes at the cost of invasive sample preparations, radiation damage to the sample, and low contrast [167].

The numerical aperture $\mathrm{NA} = n \sin\theta$ has also reached its practical limits: now, oil immersion objectives ($n \approx 1.5$) with half angular apertures of more than 60° achieve $\mathrm{NA} \approx 1.4$. Few (easy to work with) liquids have higher indices of refraction. Taking these improvements together, the smallest spot size that can be achieved is thus around 150 nm in the transverse direction and 400 nm in the axial direction.

III.1.3 Beyond the diffraction limit

Objects may be distinguished from one another at a subdiffraction scale by using a combination of methods including structured illumination, stochastic fluorophore activation, and basic data processing.

As an example of the latter, if we approximate the imaging system as a linear system, i.e., where the measured image can be obtained by applying a linear operator (convolution by the PSF) to the original sample (the emitter's original intensity distribution), it is in principle possible to mathematically invert ("deconvolve") the imaging operator to reconstruct a higher resolution image, by solving a system of linear equations. Unfortunately, theoretical results indicate that the performance of such an approach is strongly limited by noise [168, 169].

Nonetheless, in the context of microscopy, this idea was first implemented by Agard et al. [170] and may achieve a twofold improvement [171].

Furthermore, Rayleigh's criterion does not limit the ability to determine to very high accuracy the position of a single point emitter. For example, the center of a single spot can be estimated to a precision length smaller than the size of the spot itself by fitting the emission pattern to a known PSF, or an approximation of it, such as a Gaussian. The central limit theorem then suggests that the accuracy of such a calculation should be proportional to the inverse square root of the number of observed photons.

By determining the approximate position of emitters over a time series of fluorescence images, where the low density of fluorescent markers ensured their spatial separation, Morrison et al. tracked the diffusion of individual low-density receptors on cell membranes, with a resolution of $\sim 25\,\text{nm}$, well below the diffraction limit [172, 173].

Even as early as in 1995, Betzig suggested that such a localization strategy may be applicable in more densely labeled samples as well, provided that "unique optical characteristics" could be imparted on individual fluorophores [174]. Such "unique characteristics" would allow distinguishing the signals arising from each of the fluorophores; thus, the fluorophores underlying each diffraction spot could then be localized with subdiffraction accuracy [174].

Betzig's original suggestion was to discriminate certain molecules that would exhibit a random spread in their zero phonon absorption line width [174]. However, it was instead the serendipitous discovery of a photoconvertible fluorescent protein, that is, a fluorescent protein whose emission spectrum can be modified by a light-induced chemical modification [175], as well as the development of optically switchable constructs based on organic dyes [176], that provided the critical advance toward the realization of this proposal in biological samples.

Briefly, the light-induced conversion of probes to a fluorescent state at a slow enough rate ensures that only a few probes are emitting at any given time even if the sample itself is densely labeled, thus generating the sparsity needed for localization in dense environments [56–58]. Both labeling approaches were shown to be amenable to this technique: the approach based on fluorescent proteins was named (fluorescence) photoactivated localization microscopy ((F)PALM) [57, 58]; and the approach based on organic dyes, stochastic optical reconstruction microscopy (STORM) [56]).

While this review will primarily focus on techniques that rely on the stochasticity of photoconversion to temporally separate the emission of different fluorophores, it is also possible to exploit another physical phenomenon to enforce this separation. Specifically, as early as in 1994, Hell et al. noted that while the diffraction limit imposes a lower bound on the size of excitation spots, it is possible to decrease the size of this spot by "deexcitation" (stimulated emission depletion, STED) of the fluorophores located on its edges [177]. Specifically, this deexcitation is carried out by alternatively exciting fluorophores within a small region of the sample and immediately illuminating a doughnut-shaped area around this region with a

depletion laser, bringing the fluorophores back to their ground state. The intensity profile of this second region is also diffraction limited; however, given enough time, only the fluorophores close to the exact center of the doughnut (where the deexcitation intensity is zero) stay active. Measuring the fluorescence of these remaining fluorophores thus realizes a point spread function that is effectively smaller than the diffraction limit.

A similar approach, relying on the readout of fluorescence along thin stripes rather than small spots, was also developed, under the name of saturated structured-illumination microscopy (SSIM) [135]. This method relies on the observation that high spatial frequencies in the fluorophore distribution can be "brought back" to a lower frequency under illumination by a similarly high frequency pattern (i.e., by observing the beats between the two patterns) [178]. Using linear optics (structured illumination microscopy, SIM), the illumination pattern itself is diffraction-limited, and thus the resolution improvement of SIM is limited to a factor of 2 over diffraction-limited microscopy; however, the nonlinearity offered by the saturation method described above allows the generation of higher-frequency patterns and thus further gains in resolution [135].

Ultimately, the fundamental basis for any of these techniques is to note that the diffraction limit was derived under certain "standard", but not absolute, hypotheses: that all fluorophore positions must be recovered from a single image and that the signal captured depends linearly on the excitation. Attacking the first condition, by spreading the information across multiple frames, is the approach taken by stochastic photoconversion. STED and SSIM, additionally, also violate the second condition, by operating in a nonlinear regime.

The large improvement in resolution afforded by structured illumination and stochastic activation of fluorophores, together termed super-resolution microscopy, immediately opened the door to a large number of discoveries. As early as in 2007, Shroff et al. demonstrated the ability of two-color super-resolution to resolve the relative positions of pairs of proteins assembled in adhesion complexes, the attachment points between the cytoskeleton of migrating cells and their substrates, which otherwise seem entirely colocalized in diffraction limited microscopy [179].

Chapter III.2

The localization problem

The localization problem is the first step in the analysis of a super-resolution dataset and involves finding the position of a fluorescent molecule, $\mathbf{x}_0 = (x_0, y_0)$, from an image $\mathbf{I}$. The image itself is thought of as a matrix, whose elements describe individual intensities at each pixel.

In order to localize a fluorophore, we must have a model describing the expected mean number of photons per frame in pixel $\mathbf{x}$ given the fluorophore location at position $\mathbf{x}_0$, $\lambda(\mathbf{x}; \mathbf{x}_0)$. Typically, $\lambda(\mathbf{x}; \mathbf{x}_0)$ is given by the point spread function of the imaging system. The intensity at each pixel at location $\mathbf{x}$ is itself distributed randomly, according to a distribution $p(I(\mathbf{x})|\lambda(\mathbf{x}; \mathbf{x}_0))$, due to shot noise and readout noise.

We begin by describing readout noise (section III.2.1) and follow with a discussion on identifying "regions of interest" (ROIs) containing fluorophores (section III.2.2). Once positively identified, we draw from our discussion on maximum likelihood in order to describe inference frameworks used in localization in section III.2.3. While theoretically attractive, maximum likelihood methods may be computationally expensive and require good noise models to outperform simpler approaches. For this reason, we describe performance criteria of localization methods ultimately used to judge whether the computational cost of a method is warranted in section III.2.4. Sections III.2.5 and III.2.6 describe simpler localization strategies, including least-squares fit. Subsequent sections tackle generalizations of the methods discussed thus far: 3D super-resolution in section III.2.7, simultaneous fitting of multiple emitters in section III.2.8, and deconvolution-style approaches in section III.2.9. Finally, we end with a note on drift correction in super-resolution (section III.2.10) without which the best localization methods are of limited value.

III.2.1 Readout noise in single molecule experiments

Intuitively, one can expect photon shot noise to be partly responsible for reducing the accuracy of localization methods. Indeed, localization must be achieved with few photons per frame as the total photon budget of most fluorophores, meaning the number of photons collected before the fluorophore undergoes irreversible photobleaching, is limited to hundreds or thousands of photons [57, 180]. While greater brightnesses can be achieved by using quantum dots as fluorescent markers [181], they remain more challenging to deliver into cells and present toxicity concerns [182].

Perhaps more unexpectedly, accurate localization also requires a model describing how a fluorophore's emitted photons are converted into a camera readout. For instance, at a given illumination level, assuming an average number of photons strike the sample per frame per unit area, one may naively expect the camera's readout at a given pixel, $I = I(x)$, to be a Gaussian random variable identical for all pixels, or at least well approximated by such a description. In fact, as we now discuss, both Gaussian and identical assumptions are violated in practice.

Since few photons hit each camera pixel on any given frame, the Poisson limit theorem states that given the average number of photons λ for this pixel, the distribution of the actual number N_p of such photons follows a Poisson distribution ("shot noise")

$$p(N_p|\lambda) = \frac{\lambda^{N_p}}{N_p!}e^{-\lambda} \tag{III.2}$$

where for notational simplicity, we let $\lambda = \lambda(x; x_0)$.

The total noise of the measurement arises from the convolution of this shot noise by a camera readout noise, that is neither necessarily normally distributed, nor pixel-independent [183]. In other words, the readout I at a camera pixel is distributed according to a distribution $p(I|N_p)$ that is non-normal and pixel-dependent. As later described, we will use both knowledge of $p(N_p|\lambda)$ and $p(I|N_p)$ to address the localization problem [184].

III.2.1.1 Camera-specific readout

Two technologies, with different readout distributions, are widely used for single molecule imaging [184]: the older EMCCD (electron-multiplication charge coupled device), where the electrons produced by a photon hitting a pixel are collected and amplified by chip-wide electronics, and the more recent sCMOS (scientific complementary metal oxide semiconductor), which offer higher sensitivity and read rates, at the cost of pixel-to-pixel noise variation ("fixed pattern noise"), by performing signal amplification at the pixel and column level [184].

The noise distribution of an EMCCD camera follows from its amplification mechanism [185] where a photon hitting a pixel is converted into electrons. Chip-wide multiple charge-carrier multiplication (CCM) stages then amplify this electronic signal serially, one pixel at a time.

Specifically, each electron entering a stage has a low probability p of giving rise to an output of two electrons; otherwise, no amplification takes place and a single electron is output with high probability $1 - p$. Repeating this process across a large number of stages yield an exponentially distributed number of electrons arising from this single photon [185]

$$p(I|N_p = 1) \propto e^{-I/M} \tag{III.3}$$

where the multiplication factor m is itself weakly pixel-dependent, due to manufacturing imperfections. The distribution of the output from the amplification stage for N_p photons simultaneously hitting a single pixel is the N_p-fold convolution of the one-photon distribution [186, 187]

$$p(I|N_p) \propto I^{N_p - 1} e^{-I/M} . \tag{III.4}$$

After amplification, the electronic readout stage itself introduces both Gaussian noise, of standard deviation σ, which needs to be convolved to this distribution, and an offset in the number of counts ("dark count"), c_0, considered constant [186].

The other technology, sCMOS cameras, offer higher sensitivity and readout rates by attaching an individual amplification stage to each pixel. This different amplification technology yields a normally distributed readout

$$p(c|N_p) \propto \exp\left(-\frac{(c - c_0 - mN_p)^2}{2\sigma^2} \right) , \tag{III.5}$$

but the gain m, offset c_0, and variance σ^2 all vary (relatively) strongly from pixel to pixel [188].

From the distribution of camera readouts for a given number of photons, $p(I|N_p)$, and the distribution of photon counts, $p(N_p|\lambda)$, we may compute the probability distribution of the camera readout I given λ, by marginalizing over the unobservable number of photons N_p

$$p(c|\lambda) = \sum_{N_p \geq 0} p(c|N_p) p(N_p|\lambda) . \tag{III.6}$$

As we will see in the next section, this distribution is essential for our goal of estimating x_0 (on which λ depends).

As earlier mentioned, numerical estimation of this sum (which also matches experimental observations) demonstrates that $p(I|\lambda)$ is highly skewed for EMCCD cameras [186], thus violating the normally distributed noise assumption (figure III.4). In the case of sCMOS cameras,

numerical estimation of the sum in equation III.6 also yields a non-normal distribution $p(I|\lambda)$; moreover, and more importantly, this distribution changes from pixel to pixel due to the variability of m, c_0, and σ^2 [188].

In principle, one may also infer m and c_0 directly from $p(I|\lambda)$. With these two parameters at hand, and furthermore knowing, from equations III.4 and III.5, that the mean of I is a linear function of N_p, we can obtain an estimate of N_p at each pixel given I. This estimate is useful in evaluating the localization accuracy of the methods we will later explore in section III.2.6.2. However, the central quantity, moving forward, is $p(I|\lambda)$.

III.2.2 Detecting single molecules

We have previously described how the camera readout, $\mathbf{I}$, is related to the illumination level, λ, through the distribution $p(I|\lambda)$. As we will discuss in section III.2.6, physical models of spatial localization allow us to estimate, for given fluorophore parameters $\mathbf{\Theta}$, the value of λ at each pixel $\mathbf{x}$, $\lambda(\mathbf{x};\mathbf{\Theta})$. The fluorophore parameters $\mathbf{\Theta}$ minimally include the position of the fluorophore (as we had described earlier), but may also include its brightness [173], orientation [189], velocity [190], or other properties.

From $p(\mathbf{I}|\lambda)$ and $\lambda(x;\mathbf{\Theta})$ we obtain a distribution of images conditioned on $\mathbf{\Theta}$

$$p(\mathbf{I}|\mathbf{\Theta}) = \prod_{\mathbf{x}} p(\mathbf{I}(\mathbf{x})|\lambda(\mathbf{x};\mathbf{\Theta})) \tag{III.7}$$

where we have assumed that readout noise is uncorrelated across pixels.

We may, in principle, fit the entire image and simultaneously localize a large number of fluorophores. This is a difficult task, which we will address in section III.2.8. Alternatively, we may crop out ROIs centered around "emission-like" patterns, as a prelude to their further analysis [173, 191]. Mathematically, this is equivalent to marginalizing over the positions outside of the ROI, i.e., ignoring the dependence of the image within the ROI on the positions outside of it. We explain this here with the caveat that, even today, the selection of these ROIs is often treated in an ad hoc manner, with limited theoretical justification [192].

III.2.2.1 Laplacian of Gaussian filter

One may expect that ROIs could be chosen by locating pixels whose intensity go beyond some preset threshold. Such an approach cannot achieve high identification levels of relevant regions, in particular due to the presence of large amplitude and low spatial frequency background noise. Instead, a commonly used approach (and one of the few for which a theoretical basis has been offered) is to enhance features of a characteristic size σ (chosen to be that of a diffraction-limited spot) by convolution of the raw image $I(\mathbf{x}) = I(x, y)$ with a

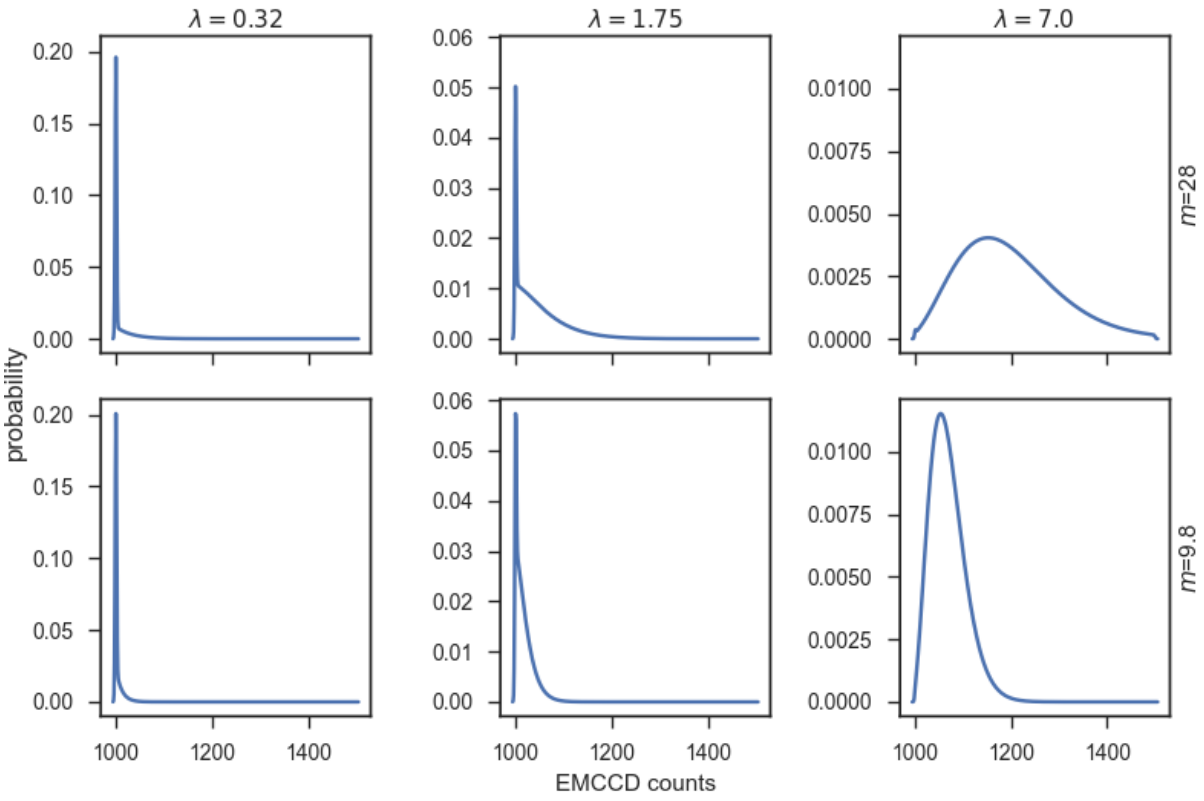


Figure III.4:
Probability densities of EMCCD camera readout counts can be highly non-Gaussian. Here, we numerically evaluated equation III.6 via a simple Python script for different mean photon numbers, λ (0.32, 1.75, 7.0) and multiplication levels m (9.8, 28). The dark count c_0 was set to 1000 and the readout noise standard deviation σ to 10.

Laplacian of Gaussian kernel $K(x, y)$ [193],

$$K(x, y) \propto \left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right] \exp\left[-\frac{x^2 + y^2}{2\sigma^2} \right]; \tag{III.8}$$

i.e., the convolved image $I'(x)$ is

$$I'(x, y) = \sum_{\delta x, \delta y} K(\delta x, \delta y) I(x - \delta x, y - \delta y) \,. \tag{III.9}$$

In this convolved image, features of a characteristic size σ have been "enhanced" to appear as sharp peaks. Peaks with a value greater than a given threshold can then be selected as originating from a single molecule and deserving further processing. This threshold is usually empirically chosen [193], for example by picking as many peaks as possible while avoiding peaks that "look like" noise (as tested during the following processing stages).

However, if we have a good model of the background noise, we can also estimate (by simulation) the distribution of peak values that would be obtained from convolving an image only constituted of background noise, and then choose a threshold value that satisfies a user-specified false-positive p-value (that is, such that the probability of observing peaks with a value greater than the threshold in a convolved pure noise image is p) [194].

Briefly, the theoretical justification for equation III.8 relies on matched filter theory [195]. Matched filter theory indicates that, if we are in the presence of additive white noise (i.e., if the differences between the observations and the true values constitute a random signal with constant spectral power density), the best linear filter to retrieve the original distribution is the convolution by the spatially reversed PSF itself ($I'(\mathbf{x}) = I(\mathbf{x}) PSF(-\mathbf{x})$). In Fourier space, such a filter corresponds to multiplication by the conjugate of the Fourier transform of the PSF. Furthermore, empirical observations establish that the spectral power density (the square of the magnitude of the Fourier transform) of background fluorescence noise, not to be confused with camera readout noise, in an image approximately follows a power-law, $\left|\mathcal{F}\{I\}(\mathbf{k})\right|^2 \propto |\mathbf{k}|^{-s}$, with $s \approx 2$ (where $\mathcal{F}\{\cdot\}$ denotes the Fourier transform) [193]. Thus, in order to apply the matched filter result, we first need to transform our data so that it exhibits white noise (whitening); this is done by multiplying the data, in Fourier space, by the filter $H(\mathbf{k}) = |\mathbf{k}|^{-s/2}$ (so that $\left|\mathcal{F}\{H \cdot I\}\right|^2 = 1$). The combination of both steps (whitening and convolution by the spatially reversed PSF) corresponds to the multiplication, in Fourier space, by the filter

$$\mathcal{F}\{K\}(\mathbf{k}) \propto |\mathbf{k}|^{-s/2} \overline{\mathcal{F}\{PSF\}(\mathbf{k})} \tag{III.10}$$

(where the overbar indicates complex conjugation). In the case where $s = 2$ and the PSF is modeled as a Gaussian, equation III.10 indeed corresponds to the Laplacian of Gaussian filter described in equation III.8 [193].

Since super-resolution datasets often contain many consecutive frames, additional improvements beyond whitening filters may be used. For instance, since background fluorescence varies slowly over time, it is possible to empirically decrease the influence of background fluorescence by working on difference images, that is, the difference in intensity between a frame and the next one [57]. The switching-on of a molecule then appears as a positive peak in the data, whereas its disappearance is a negative peak, both of which may be further selected using the whitened matched filter (equation III.8) [57].

III.2.2.2 Errors in emitter identification

The output of this initial analysis is a list of ROIs, where a single molecule is assumed to have been fluorescent. Metrics, which we now introduce, can be used to quantify the quality of the list. For any method, such metrics are typically calculated from synthetic data, where the ground truth is a priori known, which is not the case with real data. Therefore, the metrics provide only an estimate of the method result quality. If the data treated is substantially different than the synthetic data the metrics were calculated on, this estimate may be quite inaccurate.

These metrics below are expressed in terms of two kinds of possible errors: some molecules may have been missed by the detection algorithm (false negatives, FN), and some regions of interest may have mistakenly been drawn somewhere where there was, in fact, no molecule (false positives, FP) [196].

If we denote TP the number of true positives (correctly drawn regions of interest), two fundamental measures of accuracy are possible: the precision (quantifying false positives)

$$p = \frac{TP}{FP + TP} \tag{III.11}$$

and the recall (quantifying false negatives)

$$r = \frac{TP}{FN + TP}\,. \tag{III.12}$$

In order to directly rank different methods, it is convenient to combine these two measures into a single quantity. Such quantities include the Jaccard index [192, 196]

$$\mathrm{JAC} = \frac{TP}{FN + FP + TP} \tag{III.13}$$

and the F_1-score (or F-measure) [197]

$$F_1 = \frac{2}{1/p + 1/r}\,. \tag{III.14}$$

Modern localization methods are typically able to achieve high precision ($p \geq 95\,\%$) while still having limited, though widely varying, recalls ($r \approx 25\,\%$ to $75\,\%$) [192]; this latter value thus also limits the achievable Jaccard index and F_1-score.

III.2.3 Maximum likelihood localization

Having segmented our image into regions and identified whether such regions contain a single molecule, we now turn to the problem of localization within an ROI using maximum likelihood. In general, maximum likelihood estimation aims at finding the set of parameters $\boldsymbol{\Theta}$ that maximizes the likelihood of the observation, i.e., the probability of observing the actual data given the model,

$$\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} p(\mathbf{I}|\boldsymbol{\Theta})\,. \tag{III.15}$$

Such an estimator is optimal in the sense that it achieves the Cramér-Rao lower bound [198–201]; that is, it is provides an unbiased estimator (i.e., whose expected value is the true value) with a variance as low as possible.

In our case, the probability $p(\mathbf{I}|\boldsymbol{\Theta})$ is the product over each pixel $\mathbf{x}$ of the probability of observing the actual pixel value $\mathbf{I}(\mathbf{x})$, expressed as a function of the fluorophore parameters $\boldsymbol{\Theta}$ (equation III.7).

The maximization of equation III.15 can be carried using out-of-the-box numerical approaches, such as gradient descent [200]; practical implementations of such a method in a super-resolution context (which achieve the CRLB) are available for both EMCCD and sCMOS cameras [184, 202, 203].

The actual value of the CRLB (5 nm to 50 nm) depends strongly on a number of experimental parameters, most importantly the number of photons that can actually be collected [184, 202, 203].

Despite the theoretical optimality of the MLE (in the CRLB, or mean-squared error, sense), the necessarily imperfect knowledge we have about the imaging system (background fluorescence, the PSF, the camera noise) reduces its performance. In fact, PSF mis-specification or imperfections degrade the performance of the method and may even lead to overly optimistic accuracy estimations [201, 203, 204]. It thus remains useful to study simpler approaches, which can take advantage of empirical corrections.

III.2.4 Additional super-resolution performance metrics

While the Jaccard index, equation III.13, and the mean-square error of a single molecule's localization are good performance metrics, even perfect localization cannot reconstruct a biological structure that is poorly labeled [57].

To assess the quality of a reconstruction, Fourier ring correlation (FRC), a method originally developed for cryo-electron microscopy, is employed [205, 206]. Briefly, in this method, the collected single molecule events are randomly split into two datasets, which are used to create two independent reconstructions I_1 and I_2 of the structure. The "consistency" between these two reconstructions is then used as a quantification of their resolution [205, 206]. This consistency is obtained, as the name implies, by computing the Fourier transforms, $\mathcal{F}\{I_1\}(\mathbf{q})$ and $\mathcal{F}\{I_2\}(\mathbf{q})$, of the images, and computing the normalized correlation between "rings" of constant spatial frequency magnitude $|\mathbf{q}| = q$,

$$\mathrm{FRC}(q) = \frac{\sum_{|\mathbf{q}|=q} \mathcal{F}\{I_1\}(\mathbf{q})\overline{\mathcal{F}\{I_1\}(\mathbf{q})}}{\left(\sum_{|\mathbf{q}|=q} \left|\mathcal{F}\{I_1\}(\mathbf{q})\right|^2 \sum_{|\mathbf{q}|=q} \left|\mathcal{F}\{I_2\}(\mathbf{q})\right|^2\right)^{1/2}} \tag{III.16}$$

(where the overbar indicates complex conjugation).

This formula yields, for each magnitude of spatial frequency, the degree of correlation, normalized between -1 and $+1$, to which the features of that characteristic size are correlated between the two independent reconstructions. In fact, it is this separation of length scales that motivates the use of correlation in Fourier space.

For relatively large sized structures, using a random half of the events does not greatly diminish the quality of the reconstruction; thus, the two reconstructions should be highly correlated. Conversely, for structures too small to be well resolved, there is no reason to expect the two reconstructions to be highly correlated and, consequently, the FRC should be smaller.

We may then select a conventional threshold FRC (typically, $\mathrm{FRC}(q) = 1/7$) and report as "the resolution" the corresponding characteristic size beyond which the threshold is exceeded [205, 206]. Interestingly, this measure tends to indicate that nowadays, the main factor limiting the resolution of reconstructed static structures is typically the labeling density rather than the accuracy of the single molecule localization itself [205, 206].

III.2.5 Simplified localization approaches

We have seen that while maximum likelihood localization is theoretically the method that achieves the lowest mean-squared error, imperfect knowledge of the imaging system characteristics may make other localization methods preferable. Additionally, maximum

likelihood calculations are typically computationally expensive and implementations often run on specialized hardware such as graphical processing units (GPUs) [188, 203]. Thus, it remains useful to study simpler, possibly less model-dependent, approaches.

III.2.5.1 Centroid localization method

An intuitive, simple, and extremely fast approach to the localization problem is to compute the average of the pixel coordinates $\mathbf{x} = (x, y)$ within a ROI, weighted by their intensities $I(\mathbf{x})$ [191].

In such a method, it is crucially important to first subtract away any background fluorescence I_b from the ROI [207], such that the estimated localization is

$$\hat{\mathbf{x}} = \frac{\sum_{\mathbf{x}}(\mathbf{I}(\mathbf{x}) - I_b)\mathbf{x}}{\sum_{\mathbf{x}}(\mathbf{I}(\mathbf{x}) - I_b)} \tag{III.17}$$

where the sum is over the region of interest. Background subtraction is important because in its absence, the weighted average equation III.17 becomes a weighted average between the true centroid and the ROIs geometric center.

However, even with this correction, the method remains unsuitable for high-resolution localization [208]. One simple reason is that, even under the reasonable assumption that the physical PSF is symmetric (and thus its centroid should yield the fluorophore position), this is not the case for the camera readout, which is measured on a discrete pixel grid. Even worse, the centroid of the camera readout does not necessarily coincide with the centroid of the physical PSF (again due to pixelation) [208]. Still, the extreme simplicity of the method has led to its use as a minimal baseline against which other approaches can be compared [192].

III.2.5.2 Finding the point of radial symmetry

The centroid method we just described attempts to localize an event with subpixel resolution by identifying its "geometric center". Other definitions of "geometric center" have been proposed, most notably the radial symmetry approach [207, 209]. Briefly, this approach attempts to find the point that best approximates a "radial center of symmetry" for the image.

In this method, the gradient of the signal is calculated either at each pixel [210] or at each point where four adjacent pixels (or, in the 3D case [207], eight adjacent voxels) meet [209, 210]. The line defined by this point and gradient pair is taken as approximating a local axis of symmetry for the image. If all such lines were to intersect with each other at a single point, such a point would be a reasonable definition of the radial symmetry center. Because this is

not the case, the radial symmetry center is instead defined as the point that minimizes its total distance to all such lines [210], possibly with an appropriate weighting factor [207, 209].

Specifically, it is reasonable to weight lines inversely proportionately to their distance from the center of the image. Since this center is yet unknown, the weighting is instead done using the inverse distance to the centroid (as computed above) [207, 209]. Most importantly, an analytical expression can be derived to compute the radial symmetry center thus defined [207, 209]; as such, this method is extremely rapid.

While simulations indicate that this method yields high, close to CRLB-level localization accuracy of single events at a high speed [192, 209, 210], they also show that its performance degrades extremely quickly for high-density data, being unable to correctly localize events that were not well separated from the others [192].

III.2.5.3 Correlation

As discussed earlier, the good performance of the Laplacian of Gaussian kernel for event detection was justified on the basis of simple noise and PSF models (section III.2.2). We now extend this approach to tackle the localization problem itself.

In this approach, a peak's position is determined by computing the correlation between the image and the model PSF (although using the Laplacian of the PSF may work better from a theoretical point of view, as discussed above, it is the PSF itself that is typically used), and finding the position at which this correlation is maximal. The same background removal approaches as for centroid calculations may be used [208]; however, they are less important, as adding a constant background to the image simply shifts the filtered image by a constant and thus does not affect the maximum's position.

The correlation of two images is only defined for integer coordinates, so additional work is needed to obtain a subpixel localization. A simple way to do so is to fit the values of the correlation in the vicinity of the maximum with a continuous, peaked model function (e.g., a parabola) [208] and use the maximum of the latter. A more sophisticated approach is to compute this correlation after Fourier-resampling both the image and the model PSF to a higher resolution. Such a resampling is achieved by taking the Fourier transform of the image, zero-padding it to a higher spatial frequency, and taking the Fourier transform back. Correlation in real space corresponds to point-wise product in Fourier space; thus, the desired procedure amounts to computing the point-wise product of the Fourier transforms of the image and the PSF, zero-pad it, Fourier transform the padded product back into real space, and then select those coordinates at which the correlation attains its maximum [211, 212]. To sidestep the computational cost of the Fourier transforms that upsampling requires, such methods are typically first run with a limited upsampling to yield a low resolution localization and then run again with higher upsampling but only in a small neighborhood around the position selected by the first iteration [211, 212].

An important advantage of correlation-based localization is that it can be directly used for any experimentally measured PSF. For example, in particle tracking (an early application of subpixel localization [208]), one can use the image of a molecule in one frame as the model PSF for the next frame. [208] In super-resolution experiments, this approach has been suggested to analyze thick-sample data, which typically exhibits highly distorted PSFs [212] in the absence of specialized optical corrections [165]. In this case, the distorted, sample-specific PSF is measured at the beginning of the experiment by imaging a point-source at different depths; detected events are then localized by correlation with this PSF [212].

III.2.6 Least-squares fitting and model PSFs

The previous section covered methods that require limited assumptions regarding the PSF; for instance, that it be radially symmetric or invariant across the dataset. Here instead we focus on an approach, least-squares fitting, that demands no such assumptions but that does require a form of the PSF.

While, in theory, maximum likelihood achieves the optimal mean square error when an accurate PSF model is available (section III.2.3), the least-squares method is widely used [172, 173] because of the good performance of readily available, fast, and robust algorithms [192, 194]. Although we will first focus on the common case of fitting a Gaussian PSF model, we will then discuss possible corrections to this model.

III.2.6.1 Gaussian PSF least-squares fitting

Since the theory of least-squares fitting, as with maximum likelihood (section III.2.3), can be described independently of the model PSF's exact functional form, we will, for simplicity, assume a Gaussian PSF. This choice is one of the earliest in use, offers mathematical simplicity, and maintains good performance.

Specifically, we model the image $\mathbf{I}_0$ arising from a fluorophore as a two-dimensional Gaussian,

$$\mathbf{I}_0(\mathbf{x}; A, \mathbf{x}_0, \sigma, I_b) = I_0 + A \exp\left(-\frac{|\mathbf{x} - \mathbf{x}_0|^2}{2\sigma^2}\right). \tag{III.18}$$

The unknown amplitude A, center $\mathbf{x}_0$ and standard deviation σ, as well as the unknown, locally constant mean background I_b, are parameters collectively regrouped as $\mathbf{\Theta}$, the fluorophore characteristics, that we now want to infer. Furthermore, despite the subtraction of the background I_b, the measured image still differs from the model by a noise term of mean zero. Next, it is also possible to assume that some parameters are a priori known such as σ or I_b, for example, they may be independently estimated from the image intensity far away from

the fluorophore [213]. It is also possible to improve this model by averaging the PSF over each pixel [56].

One may then infer the remaining set of parameters minimizing the sum of squared differences between the observed intensity and model provided by equation III.18, weighted by the signal variance at each pixel. Numerically, this is a classic least-squares minimization, for which fast and robust implementations, such as the Levenberg-Marquardt algorithm [183], are available.

The maximum likelihood framework (section III.2.3) and least-squares fitting are identical, even for non-normal PSFs, if the noise at each pixel is assumed to be independent and drawn from the same normal distribution with unspecified variance

$$l \propto -\sum_{\mathbf{x}} (\mathbf{I}(\mathbf{x} - \mathbf{I}_0(\mathbf{x}; \mathbf{\Theta})))^2 + \text{constant} \tag{III.19}$$

where l denotes the log-likelihood.

III.2.6.2 Least-squares fitting localization accuracy

Thompson et al. provided a theoretical analysis of least-squares fitting accuracy in the presence of normally distributed background noise as well as photon counting noise (section III.2.1), as a function of the PSF's standard deviation (the "spot size") s, the pixel size a, the number of photons in the event N_p and the standard deviation of the background noise b (figure III.5) [214].

For simplicity, we limit ourselves to re-deriving Thompson's results in the case of a 1D model and assume that for a fluorophore at position x: (i) the expected number of photons at the ith pixel (i.e., the PSF model) is $N_i(x)$; (ii) the variance is $\sigma_i^2 = Ni(x) + b^2$ (i.e., the sum, in quadrature, of the photon counting noise, $N_i(x)$, and the background noise, b^2); and (iii) the detected photon number at that same pixel is y_i. By definition, the fitted position, $\hat{x}$, is obtained by minimizing the weighted sum of square residuals, i.e.,

$$\frac{\partial}{\partial x} \sum_i \frac{(y_i - N_i(x))^2}{\sigma_i^2}\bigg|_{x=\hat{x}} = 0\,. \tag{III.20}$$

By expanding $N_i(x)$ in the above to first order in $\hat{x}$ around the true underlying position x_0 ($N_i(\hat{x}) \approx N_i(x_0) + N_i'(x_0)(\hat{x} - x_0)$) and solving for $\Delta x = \hat{x} - x_0$, we directly derive the mean square error of the fitted center's position

$$\left\langle (\Delta x)^2 \right\rangle = \frac{1}{\sum_i (N_i'^2/\sigma_i^2)}\,. \tag{III.21}$$

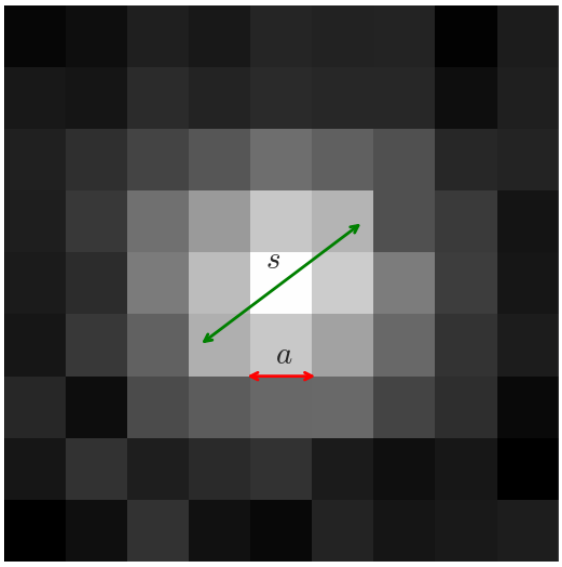


Figure III.5:
The image of a point source by a microscope can be approximated as a Gaussian of standard deviation s. Collecting this image on a camera further pixelates it with pixel size a. The noisy intensity profile was generated using a simple Python script.

While this sum can be evaluated numerically, we can also simplify it under reasonable approximations. We ignore, for now, the effects of pixelation ($a \to 0$). In this case, under a Gaussian PSF model, the expected number of photons at pixel i is $N_i = (N_p/\sqrt{2\pi}\sigma)\exp(-i^2/2\sigma^2)$, and the sum in equation III.21 can be replaced by an integral.

In general, this integral is not analytically tractable but it can be asymptotically evaluated in two limits: (i) dominant photon-counting noise ($N_i \gg b^2$, so $\sigma_i^2 \approx N_i$) and (ii) dominant background noise ($N_i \ll b^2$, so $\sigma_i^2 \approx b^2$). These two cases respectively yield

$$\left\langle (\Delta x)^2 \right\rangle_1 = \frac{s^2}{N_p} \quad \text{and} \quad \left\langle (\Delta x)^2 \right\rangle_2 = \frac{8\pi s^4 b^2}{a^2 N_p^2}\,. \tag{III.22}$$

Since each expression dominates the other in the limit where it has been derived, the authors suggested the following interpolation formula [214]:

$$\left\langle (\Delta x)^2 \right\rangle = \frac{s^2}{N_p} + \frac{8\pi s^4 b^2}{a^2 N_p^2}\,. \tag{III.23}$$

The pixelation noise's main effect (arising from a nonzero a) is to increase the photon counting noise term $\left\langle (\Delta x)^2 \right\rangle_1$. Specifically, the PSF's spatial variance, s^2, appearing in this term should be increased by the spatial variance of a square pixel of size a, which is $a^2/12$. [214] The final expression for the uncertainty of Gaussian fitting is thus

$$\left\langle (\Delta x)^2 \right\rangle = \frac{s^2}{N_p} + \frac{a^2}{12N_p} + \frac{8\pi s^4 b^2}{a^2 N_p^2}\,. \tag{III.24}$$

Although equation III.24 is widely used to report localization accuracies [57, 58, 215], the summation in equation III.21 can also be evaluated numerically. [214] This numerical estimate indicates that equation III.24 actually overestimates the localization accuracy (in the relevant regime of parameters) by approximately 10 %. [214] This is a discrepancy that has also been reported from experimental comparisons [56].

An interesting consequence of equation III.21 is that the mean square error is minimal for a nonzero pixel size a ($\partial(\Delta x)^2/\partial a = 0$). In other words, it is counter-productive to make the pixel size as small as possible. Instead, its optimal size is close to the spot size s. Intuitively, this is due to the compromise between the higher spatial information gained from each pixel when the pixels are smaller and the averaging out of background noise when the pixels are larger.

In practice, Gaussian PSF fitting has been shown to achieve nanometer-resolution. For example, Yildiz et al. have used this approach to show that the motion of fluorescently labeled myosin V enzymes along their tracks occurs in steps of variable size that can be grouped in consecutive pairs whose sizes add up to 74 nm (fluorescence imaging with one-nanometer accuracy, FIONA) [215].

III.2.6.3 Applicability of least-squares to non-normal noise

While the assumption of identically and normally distributed noise is reasonable in many applications of least-squares fitting, which is the source of its versatility and the reason many efficient algorithms have been developed, it is clearly violated in super-resolution, as described in section III.2.1.

Although many super-resolution analysis discount non-normal noise, here we discuss a variance-stabilizing transformation [194] that mitigates the effect of ignoring non-normal noise.

For simplicity, we consider only the effect of Poisson (shot) noise, whose variance is equal to its mean. Since the variance of the noise changes across the fitted ROI, the assumption of identical noise distribution is violated.

In order to correct for this non-uniformity, we exploit the following (numerical) observation, known as the Anscombe transform: if X is Poisson-distributed with both mean and variance

equal to $m \geq 4$, then $2\sqrt{X + 3/8}$ is approximately normally distributed with mean $2\sqrt{m + 3/8} - 1/4\sqrt{m}$ and, more importantly, unit variance [216]. Thus, applying this transformation to an image corrupted by Poisson noise yields an image with (approximately) uniform Gaussian noise and the classical least-squares algorithm may then be applied. Of course, the fit should not be done using the original PSF model but, likewise, the Anscombe-transformed model [194].

Note that this correction assumes that the image data is correctly expressed in units of photon counts, which requires a calibration of the readout-to-photons conversion factor as discussed in section III.2.1. Additionally, more sophisticated transforms (e.g., the generalized Anscombe transform [217]) may be used to handle more realistic non-Poisson noise models.

While, to our knowledge, the effect of a variance-stabilizing transformation for the accuracy of least-squares fitting has not been evaluated independently of other improvements, the SimpleSTORM package, which relies on it as a preprocessing step before least-squares fitting [194], was shown to exhibit a relatively strong performance [192, 194].

III.2.6.4 Corrections to the point spread function

Although we have mentioned, and it is widely quoted [191, 214], that the diffraction pattern of a point source is an Airy disc (section III.1.2), and chose to approximate this pattern with a Gaussian peak both for maximum likelihood and for least-squares fitting, we now revisit this claim.

When imaging using a high-NA objective, as commonly done in super-resolution applications, the PSF of a freely rotating fluorophore, directly derived from first-principles, is in fact closer to a Gaussian function than to an Airy function [218] thus justifying, a posteriori, the use of Gaussians for least-squares fitting.

A rotationally constrained fluorophore, which may occur, or conversely be avoided, e.g., due to the labeling strategy used [189, 219], presents additional complications. Such a constraint breaks radial symmetry, in which case the PSF may present two "lobes" [189, 218]. If a rotationally free model, such as a Gaussian, is used to fit datasets lacking radial symmetry, simulations indicate that maximum likelihood estimation can lead to substantial errors (dozens of nanometers), in particular in the case of defocused molecules (e.g., for 3D measurements) [218, 220]. Conversely, orientational information may be derived from properly fitting the observed PSF to a model PSF for anisotropic emission [189].

In the opposite extreme, highly mobile fluorophores, which move by a significant fraction of a pixel size during the time it takes to acquire a single frame [190], may distort the effective molecular PSF, which is now a weighted average of the PSF at each position visited by the molecule. Once more, ignoring this distortion leads to poor localization accuracy, whereas using a PSF model that takes motion into account not only restores the original localization

accuracy but also provides information on the instantaneous molecular velocity [190] and additional information on motion models.

We end with a note on the non-uniformity of a sample's refraction index, which introduces additional PSF aberrations, especially for thick samples [165]. This effect is has so far been treated experimentally by using adaptive optics (e.g., deformable mirrors) to properly shape the PSF [165].

III.2.7 3D localization

III.2.7.1 Cylindrical lens 3D

While our discussion, so far, has been limited to localizing single molecules in a 2D plane, most biological samples are three-dimensional and, as a consequence, there is considerable interest in obtaining volumetric fluorescence data.

In classical microscopy, this can be achieved by selectively exciting, and thus collecting, fluorescence from a single plane (multiphoton microscopy [221] or selective plane illumination microscopy (SPIM) [222]). However, such techniques remain essentially limited by diffraction. Instead, true 3D super-resolution can be achieved by encoding information about the depth of a molecule in its PSF.

Fundamentally, the techniques we have discussed up until now fit a PSF that encodes lateral but not vertical information. In other words, in 2D, the value of the PSF measured by the camera at position (x, y) when the emitter is at position (x_0, y_0) depends only on the distance between the two positions, i.e., $PSF = PSF(x - x_0, y - y_0)$. In 3D, the dependence on the true position z_0 cannot be expressed in terms of translation and the PSF would need to be of the form $PSF(x - x_0, y - y_0, z_0)$.

As early as in 1994, Kao et al. introduced a cylindrical lens in the optical path of their particle tracking setup and observed a depth-dependent PSF [223]. This depth-dependent PSF progressively switches from being a vertically oriented ellipse for molecules above the focal plane to a horizontally oriented one on the other side. Thus, the lengths of the PSF's two axes, w_x and w_y, could be estimated and converted to a depth value using a calibration table. Specifically, the relative difference between the two widths, defined as $R = (w_y - w_x)/(w_x + w_y)$, was matched with a calibration curves $R^{\text{cal}}(z)$ in order to read out the depth z while the (x, y) position was obtained by least-squares fitting to a parabolic PSF (section III.2.6).

The cylindrical lens approach was adapted for super-resolution by Huang et al. [224] who took advantage of the advent of more general nonlinear fitting procedures, allowing the determination of w_x and w_y by least-squares fit along with the in-plane position. That is, the model PSF was chosen as a Gaussian with the following parameters that need to be fitted: the position of the center and the amplitude of the PSF, similarly to the two-dimensional case along

with the PSF width and height w_x and w_y treated as independent parameters. The authors found, purely empirically, that the fitted w_x and w_y could be accurately mapped back to the molecule depth z via the use of a calibration curve $(w_x^{\text{cal}}(z), w_y^{\text{cal}}(z))$ obtained by measuring the PSF of point sources positioned at different depths, as follows: the depth z is chosen to minimize the Euclidean distance between the $(w_x^{1/2}, w_y^{1/2})$ point and the $(w_x^{\text{cal}}(z)1/2, w_y^{\text{cal}}(z)^{1/2})$ curve.

Instead of estimating depth based on ellipticity calibration curves, we may immediately adapt all methods described for 2D localization to the 3D case by simply including the depth z in the set of parameters $\boldsymbol{\Theta}$ [212, 225]. All theoretical results regarding such methods, such as the CRLB accuracy limit (section III.2.3), are then applicable. For example, we demonstrated earlier that in the presence of a highly distorted, but experimentally well characterized PSFs, the position of the maximum in the ROI's correlation with the PSF generated good localization performance (section III.2.5.3). This method is, in fact, especially applicable to 3D imaging of thick samples, as the PSF of events localized deep into the cell can be distorted by severe optical aberrations [212].

III.2.7.2 Other approaches for encoding depth information in the PSF

While the cylindrical lens approach is relatively simple from an experimental viewpoint, it only requires introducing a cylindrical lens in the optical path, it encodes depth information at the cost of lateral resolution, as it distorts the PSF. Additionally, as discussed earlier in section III.2.6.4, other phenomena can lead to elliptical distortion of the PSF, leading to spurious apparent changes in depth. Hence, additional ways to encode depth information have been proposed [225].

For example, the biplane-PALM approach relies on simultaneously imaging two planes, a few hundred nanometers from each other on the same camera [225]. This can be achieved by imaging on one-half of the camera chip the "standard" focused image and, on the other half, a slightly defocused image, obtained by splitting the collected light and reprojecting it onto the camera after a longer light path. A ROI corresponding to a single event now coincides with a pair of spots, one on each plane, that may once more be fitted by least-squares either to an experimental PSF, also measured over the two planes, or a theoretically derived one [225]. As a fluorophore is displaced along the z axis, it does not get defocused to the same degree in the two planes; this difference in defocusing thus encodes the depth information. In its first implementation, a depth resolution of 75 nm was achieved [225].

Additional z-resolution can be provided by more sophisticated procedures. For example, a spatial light modulator can be used to shape the 3D PSF into a double-helix, such that individual events are now observed as pairs of close peaks, whose relative position encode depth information [139]. This technique, to which all the previous fitting discussions apply, exhibits

a low theoretical maximal resolution (CRLB) of approximately 15 nm [139]. Interferometric PALM (iPALM) provides an even more sophisticated procedure to encode depth information [226] in which the measured image is split over three cameras, each of which measure an interference pattern between two images that are phase-shifted with respect to one another. The relative intensities of a same peak across all three cameras allow the experimenter to compute this phase shift and thus infer the event depth, with an experimentally demonstrated resolution of approximately 10 nm [226].

III.2.8 Simultaneous localization of multiple molecules

The fundamental breakthrough from which super-resolution microscopy emerged, namely achieving temporal separation of events too close to be resolved spatially, is also an important limitation. As described so far, a super-resolution acquisition scheme must ensure that only a few molecules are activated per frame, thus imposing lengthy acquisition times for densely labeled samples.

However, just as we have described various ways in which the coordinates of a single molecule can be retrieved if a model PSF is known, we could, in theory, write down an emission model for two, or more, close molecules with overlapping PSFs (given their coordinates x_0 and x_1), and then fit a ROI to such a model. This approach was pioneered by astronomers who were interested in separating images of stars in "crowded fields" (e.g., stars in distant galaxies, which appear very close to each other) and have since long ago developed such algorithms [227]. One of these algorithms, DAOPHOT (Dominion Astrophysical Observatory photometry) [227], was directly adapted for super-resolution microscopy, under the name of DAOSTORM [166].

There are a few difficulties that are associated with the simultaneous fitting of multiple molecules at a time. The first is computational; the greater the molecules simultaneously fit, the greater the number of parameters, rendering the least-squares or maximum likelihood optimization more challenging numerically.

Fortunately, it is clear that even when PSFs are slightly overlapping, it remains acceptable to cut the image into smaller regions, that are approximately statistically independent from each other, and fit them one at a time. This approximation was used by another super-resolution package developed at the same time, MFA (multiple-emitter fitting analysis) [228]. More accurately, one can also draw such regions to be bounded by areas of the image where the intensity is relatively low and are thus unlikely to contain a molecule (the approach of DAOPHOT/DAOSTORM). Again, in such cases, the problem of fitting PSFs in a region becomes independent from the fitting in another region [166], in a manner similar to how we drew ROIs for single-emitter fitting but this time with multiple fluorophores per ROI.

More importantly, simultaneous fitting of many fluorophores also presents a model selection problem: allowing for more fluorophores always result in a better (or at least, not worse) fit of a collection of spots (either the fitting algorithm can exploit the additional degrees of freedom to eliminate some residuals of the fit or, at worst, it can always set the brightness of the additional fluorophores to a very small value, thus not worsening the fit). Thus, additional criteria are necessary to prevent overfitting.

While a review of general model selection methods are outside of the scope of this work (see for example [44, section 3.3.4]), here we present two more model selection strategies specifically adapted to the problem of multi-emitter fitting used by DAOSTORM [166] and by MFA [228].

DAOSTORM first uses a peak detection algorithm (such as the one discussed in section III.2.2.1) in order to find candidate regions that may correspond to a molecule. This set of candidates is then fit, by MLE, to a multi-emitter model. The residuals of the fit (i.e., the difference between the original image and the one that a set of fluorophores at positions given by the fit would yield) are then iteratively reinserted into the original peak detection algorithm [166]. Thus, it is the sensitivity of the peak detection algorithm that provides a stopping criterion against the addition of extraneous fluorophores to the fit.

Model selection by MFA [228] relies instead on computing the log-likelihood ratio, LLR:

$$\mathrm{LLR} = -2\log\left[\frac{L(\{(x_i,y_i)\}_{\mathrm{MLE}}|\mathbf{I};\mathrm{noise})}{L(\mathbf{I}|\mathbf{I};\mathrm{noise})}\right]. \tag{III.25}$$

The numerator, $L(\{(x_i,y_i)\}_{\mathrm{MLE}}|\mathbf{I};\mathrm{noise})$, is the likelihood of the estimates given the image, assuming that each pixel's signal is independently obtained from a Poisson-distributed source with mean equal to the sum of the PSFs at this pixel (equation III.7). The denominator, $L(\mathbf{I}|\mathbf{I};\mathrm{noise})$ is the maximum value that the above-mentioned likelihood could ever attain, which it does in the case where the expected mean intensity at each pixel matches the actually observed intensity. In other words, it is the product over the pixels of the probability of observing the actual camera output if the mean expected intensity at that pixel was set to be equal to that output.

Having evaluated the "goodness" of each model (as measured by its LLR), we now need to estimate, for each model, how well the model matches the data, as compared to how well it would match random datasets generated from the model itself. Such a comparison penalizes overfitting, as the marginal improvement to the LLR, for each additional parameter, decreases sharply once the "correct" number of parameters is reached, whereas such a transition does not occur for random datasets.

More specifically, we need to estimate the probability p that the LLR of a dataset generated from the model be lower than the LLR of the real data. In other words, we need the value of the cumulative distribution for the LLR, evaluated at the LLR of the real data. According to

Wilks' theorem, this distribution can be approximated by a χ^2 distribution with a number of degrees of freedom equal to the difference between the number of pixels and the number of fitting parameters [229]. We thus obtain p simply by evaluating the cumulative distribution function of the above-mentioned χ^2 distribution [228]. Having done so for each of the models in contention, the model with the highest such probability is then selected.

III.2.9 Deconvolution-based super-resolution

We have so far focused on reconstructing coordinates of each single event with subdiffraction accuracy. However, subdiffraction imaging may be achieved by other means. For example, deconvolution microscopy achieves a twofold improvement over diffraction-limited microscopy by approximating the inverse (in a linear operator sense) of the "imaging operator", i.e., the operator that convolves a distribution of point emitters by the imaging system's PSF [170].

Here, we discuss adaptations of deconvolution-style approaches to datasets collected using single molecule localization-style techniques where additional information is encoded in temporal fluctuations of the fluorescence (i.e., stochastic switching of the fluorophores).

III.2.9.1 Compressed sensing

Contrary to localization methods discussed thus far, here we do not initially attempt to reconstruct a list of molecular positions. Instead, we want to reconstruct a higher resolution image than the one from which we started.

More specifically, we seek a "fluorophore density map" on a discrete grid $\mathbf{s}$, where each "pixel" on the grid may be smaller than the raw image, $\mathbf{I}$, physical pixels. Instead of considering $\mathbf{s}$ and $\mathbf{I}$ as matrices, we will consider them as vectors of entries (for example, by concatenating the physical columns of pixels in the image), respectively of size N and n. In this formalism, convolution by the PSF, which is a linear operator, can be understood as multiplication by a matrix $\mathbf{A}$, of size (N, n),

$$\mathbf{I} = \mathbf{A} \cdot \mathbf{s} \,. \tag{III.26}$$

Each row of the matrix $\mathbf{A}$ corresponds to a possible fluorophore position and each entry in the row corresponds to a physical pixel indicating how much a fluorophore at the row-encoded position would increase the intensity at that physical pixel.

Localization methods discussed so far correspond approximately to a setup where we know (or have a good model of) $\mathbf{A}$ (i.e., how much a fluorophore at any position affects the intensity measured at any position—in other words, the PSF) and seek to obtain $\mathbf{s}$ (i.e., the

fluorophore density map). We will focus on the same formulation first. However, we will later see that we can also attempt to recover **A** and **s** simultaneously.

The usual caveats of deconvolution microscopy, namely the sensitivity of **s** to noise and to inaccurate knowledge of **A** still apply. Moreover, as there are many more (discretized) fluorophore positions than image pixels ($N \gg n$), the problem is underdetermined. However, in the context of a super-resolution dataset, we have the additional information that we expect only a few fluorescent proteins to be "on" on each frame; that is, we have a sparsity prior on **s** (we expect most of its entries to be zero).

This class of problems (searching for approximate and sparse solutions to an underdetermined linear system) is known as compressed sensing and is well described in the mathematical literature [230]. For example, Zhu et al. showed that in the presence of photon counting noise, a solution can be obtained by searching for the vector **s** with minimum l^1 norm (i.e., sum of absolute values of components) among all those for which the l^2 norm of the residual vector, $\mathbf{I} - \mathbf{A} \cdot \mathbf{s}$ (i.e., sum of squared errors), is no larger than a noise-level dependent threshold [231]. Such a vector can then be found using standard algorithms [232].

Such a deconvolution yields, for each frame of the image stack, a sparse list of discretized molecular positions. All such lists can then be merged together to obtain a final list of molecular positions. Although the original implementation of this idea [231] yielded a relatively poor localization accuracy of $\sim 60\,\text{nm}$, it was able to recover highly overlapping events, i.e., it allowed a very fast imaging rate (6 to 15-fold faster than for single-event fitting, 2 to 3-fold faster than for a multi-emitter fitting such as DAOSTORM).

III.2.9.2 Exploiting fluorophore temporal fluctuations

Instead of using an experimental protocol designed to achieve temporally sparse photoactivation of fluorophores, it is also possible to rely on the natural blinking and bleaching of fluorophores, that occurs (to varying degrees depending on the fluorophore) even under continuous illumination.

For example, a simple way to exploit the fluorophore blinking is to compute the difference between consecutive frames of a regular fluorescence movie. In these difference images, the spontaneous switching-on of a fluorophore appears as a positive peak, whereas turning-off events, or photobleaching, appear as a negative one. Standard localization algorithms, such as PSF fitting, can then be directly applied to such images, yielding a super-resolution approach that does not require the use of photoconvertible markers (bleaching/blinking-assisted localization microscopy, BaLM) [233]. More interestingly, it is possible to exploit the fact that these temporal fluctuations in fluorescence intensity are uncorrelated between molecules (as each fluorophore undergoes stochastic switching independently from the others).

Dertinger et al. noticed that due to this stochasticity, pixels where the emission of two blinking fluorophores (quantum dots, in their case) overlap exhibit lower temporal coherence than pixels which capture the emission of a single quantum dot [234]. This observation yields a simple and elegant method, named super-resolution optical fluctuation imaging (SOFI), to obtain a superresolved image $\mathbf{I}$ [234]. At each pixel, one simply plots the value of the temporal correlation of this pixel's signal for a well-chosen time lag τ

$$I_\tau(x,y) = \langle I(x,y,t)I(x,y,t+\tau)\rangle_t \tag{III.27}$$

where $\langle\cdot\rangle_t$ denotes an average over time.

Lidke et al. exploited temporal fluctuations in order to generalize the model proposed in equation III.26 [181]. Remember that we originally wrote $\mathbf{I} - \mathbf{A}\cdot\mathbf{s}$, where $\mathbf{I}$ is the image (a size n vector), $\mathbf{s}$ the discretized fluorophore density (a size N vector), and $\mathbf{A}$ the imaging operator (an N by n matrix). In this generalization, the time dependency, over T frames, of $\mathbf{I}$ and $\mathbf{s}$, was also taken into account; these two variables are now matrices respectively of size (n,T) and (N,T), where each row encodes respectively the time-varying image intensity at a pixel and the time-varying active fluorophore density at a position. The shape of $\mathbf{A}$ is unchanged, and we still have $\mathbf{I} - \mathbf{A}\cdot\mathbf{s}$. However, in this analysis, we will also consider the imaging operator $\mathbf{A}$ as an unknown.

The problem may thus appear severely under-determined, as we are trying to reconstruct $N\times T + n\times N$ parameters ($\mathbf{A}$ and $\mathbf{s}$) while having only $n\times T$ measurements ($\mathbf{I}$). However, we can exploit the fact that in our target reconstruction, each row of $\mathbf{s}$ should represent the time-varying intensity of a single fluorophore at a fixed position; conversely, a reconstruction will be poor if some rows of $\mathbf{s}$ encode the time-varying mixture of the intensities of multiple fluorophores. From the central limit theorem, the values of $\mathbf{s}$ adopted by a mixture are necessarily "more normally distributed" than intensity values of a single fluorophore. In other words, a weighted sum of multiple iid random variables is more normally distributed than each individual variable. Thus, we can restate our objective as follows: we seek the solution of equation III.26 for which the rows of $\mathbf{s}$ are "as non-normally distributed as possible". In order to quantify the "non-normality" of the distribution of values a row of $\mathbf{s}$ takes, we compute the entropy of the distribution, $H = -\sum_{\mathbf{s}} p(\mathbf{s})\log p(\mathbf{s})$. Because the normal distribution has the maximal entropy among all distributions for a given mean and variance, we thus seek the solution $(\mathbf{A},\mathbf{s})$ of equation III.26 for which the total entropy of $\mathbf{s}$ (the sum of the entropies of each row of $\mathbf{s}$) is minimal [181]. This minimization problem is known as independent component analysis, and can be solved using the standard FastICA algorithm [235].

The outputs of this analysis are both matrices $\mathbf{A}$ and $\mathbf{s}$ with $\mathbf{A}$ giving the PSF associated with each of the fluorophores while $\mathbf{s}$ indicates, for each fluorophore, the frames on which it is active. This analysis does not directly yield superresolved coordinates; it simply separates

the PSFs of each fluorophores (into columns of **A**) starting from a dataset where they were spatially and temporally overlapping. Each of these PSFs can then be fit to obtain a superresolved coordinate for each fluorophore using any of the methods we have discussed so far [181].

III.2.9.3 Bayesian deconvolution approach for fluorescence time series

Bayesian methods may also be used to obtain both spatial (**A**, following our earlier notation) and spatiotemporal (**s**) information. For example, Cox et al. simultaneously fitted the full set of fluorophore positions, the state histories (bright, transiently dark, or irreversibly photobleached) for each fluorophore, as well as the transitions rates between these states (Bayesian analysis of blinking and bleaching, 3B) [236].

This Bayesian formulation can be seen as another approach to tackling equation III.26: instead of reducing the problem to independent component analysis, the time-evolution of the fluorophores (i.e., each row of **s**) is modeled as a Markov chain between the three abovementioned states. The true underlying fluorophore spatial distribution is then selected as the one maximizing the likelihood that the entire image stack arises from that distribution. This likelihood is computed by integrating over all possible temporal evolutions (which is done using the forward algorithm) [183, 236]. Instead of yielding a maximum likelihood estimate, one can also sample (by Markov chain Monte Carlo) spatial fluorophore distributions from the posterior derived from this likelihood [183, 236], thus yielding a super-resolved image where the intensity at each position encodes the confidence level about the presence or absence of a fluorophore there. This method is extremely demanding computationally, to the point that cloud-based implementations have been developed [237].

III.2.9.4 Richardson-Lucy deconvolution for fluorescence time series

Mukamel et al. also proposed a simpler deconvolution method (deconSTORM) taking temporal correlations into account [238]. Again, the final output of such a method is not a list of coordinates, but simply an image with a higher resolution. Specifically, Mukamel et al. based their work on Richardson-Lucy deconvolution.

Briefly, Richardson-Lucy deconvolution is an iterative approach, whereby the estimated deconvolved image $\hat{\mathbf{I}}_k$ at iteration k is derived from the estimate at the previous iteration $\hat{\mathbf{I}}_{k-1}$, as well as the measured image **I**, the PSF (assumed known, under a Poisson noise model), and a prior distribution on the true image $p(\cdot)$, which is, in the classical form of the algorithm, kept constant throughout iterations:

$$\hat{\mathbf{I}}_k = RL(\hat{\mathbf{I}}_{k-1}, \mathbf{I}, h, p(\cdot)) \,. \tag{III.28}$$

Mukamel et al. proposed to deconvolve a time series of images by running the iterations of Richardson-Lucy deconvolution in parallel; that is, at each iteration, a new deconvolution of each frame is computed. More importantly, instead of keeping the same image prior throughout the iterations, they used a different prior for each frame and updated this prior at each iteration:

$$\hat{\mathbf{I}}_k = RL(\hat{\mathbf{I}}_{k-1}, \mathbf{I}, h, p_{\mathbf{I}_k(t-1)}(\cdot)) \,. \tag{III.29}$$

Specifically, when running an iteration, the prior for the frame at time t was chosen so that the a priori probability of observing a bright pixel at a given position in that frame is increased whenever the same pixel was also bright in an earlier frame (already reconstructed during this iteration); the closer (temporally) this frame was, the stronger the contribution to the prior. In other words, each frame is deconvolved with a series of priors that, at each iteration, favors a reconstruction similar to reconstructions of the preceding frames [238].

III.2.9.5 Recovering molecular localizations from deconvolution-style approaches

We have presented deconvolution-style approaches to obtain a superresolved image without first localizing single molecules. But both methods, deconvolution and localization, may be used in tandem. That is, an initial deconvolution step identifying candidate single molecule positions (from local maxima of the deconvolved image), may be used as initial guesses in a subsequent localization step. Such an approach was implemented in the FALCON algorithm [239] that may be understood as a variant of multi-emitter fitting (section III.2.8), where the model selection step (finding the correct number of fluorophores to fit) is accomplished by an initial deconvolution.

III.2.10 Drift correction

In the absence of active correction, different microscope components drift by dozens of nanometers relative to each other during the acquisition of a single molecule localization dataset [56, 57]. This drift affects positions of measured events.

Thus, in order to combine all the localization events obtained in that dataset into a single high-resolution image, it is necessary to either (i) actively correct for this drift by measuring it in real time and displacing the sample in a compensatory manner or (ii) to estimate the drift in order to subtract it from the fitted positions.

In practice, the second option (drift estimation and subtraction) is typically chosen, as it is a purely mathematical operation, that does not require any modification to the instrument itself. In order to do so, we may track bright fiducial markers (e.g., gold nanoparticles or

fluorescent beads) on the coverslip. This can be achieved by using the same localization algorithms as used for "real" events [56, 57]. As the fiducial marker concentration can be chosen to be very sparse, tracking markers from one frame to the next is straightforward. Additionally, the high level of brightness of these fiducially ensures that they are at least as well, and typically better, localized than the events themselves, i.e., they are not a limiting factor for localization accuracy.

Since fiducials are typically bound to the coverslip, such that their motion relative to the camera matches the sample drift relative to the camera, they are less suitable for thick-sample 3D single molecule localization microscopy. From the instrumentation point of view, the use of fiducial markers in a thick sample dataset requires repeatedly switching between the imaging planes and the fiducial (i.e., coverslip) plane [212]. To avoid the need for such a movement, which complicates the experimental setup and may lead to additional drift itself, one may abandon the use of fiducials and instead rely on correlating event time-slices. In this approach, groups of events are formed by stacking consecutive frames until reaching a set number of events. The cross-correlation between event positions in one group and those in the next then exhibits a peak at a position that encodes the average displacement of the events between the two groups—in other words, the sample drift—as long as the reasonable assumption that both groups are randomly sampled from the entire structure holds [240, 241].

Neither of these methods can correct drift that occurs on the same time scale as the frame rate as drift estimation requires averaging over a large number of events. In order to increase the rate at which drift information is collected, McGorty et al. proposed instead to use a correlation drift estimator on the bright field image itself (that is, the drift is estimated by finding the shift that maximizes the correlation between a bright field frame and the previous bright field frame) [242]. Of course, it is not possible to simultaneously collect a bright field image in the visible wavelength and single molecule fluorescence in the same wavelength, as the former would swamp the latter; McGorty et al. thus collected the bright field image in the infrared spectrum. Such an approach allowed them to achieve real-time drift correction, with a 10 nm in-plane and 20 nm axial stability, at rates of a few hertz and over minutes of acquisition [242].

Chapter III.3

The counting problem

Whether two fluorescent events occurring in close spatial and temporal proximity actually come from the same fluorescent protein is an important question that was raised since early PALM experiments [57, 58].

One reason motivating this question is technical: "stacking" multiple frames together, by summing their intensities before fitting them, ensures that all photons associated with a single labeled protein are used, thereby improving the reconstructed image's final quality [57].

The question was less relevant in STORM, which relies on blinking organic dyes, as multiple dye-labeled antibodies typically bind to the same target [56]. Each labeled site is thus associated with dozens or hundreds of fluorescence events. Thus, proper assignment of each event to its original label is essentially impossible [243].

By contrast, in PALM, given a number of fluorescent events arising from a single diffraction limited spot, it was reasonable to ask whether one may enumerate the proteins, or, alternatively, quantify the protein density, that gave rise to the fluorescent signal. Furthermore, if the number of fluorescent events originated from a group of proteins that formed a complex, with each subunit individually labeled, it may then be possible to quantify protein complex stoichiometry.

Inferring protein complex stoichiometry in vivo is an important problem. Many protein complexes involved in essential cellular tasks contain multiple copies of various proteins. For example, *E. coli*'s flagellar motor is composed of dozens of proteins all appearing in dozens of copies [244].

What is more, protein complex stoichiometry may well be dynamical since a pool of freely diffusing protein subunits available to a protein complex changes over time [245, 246]. The FliM bacterial flagellar switch protein [246] is a typical example.

Furthermore, determining a complex's stoichiometry can also help understand a complex's operation. For example, asymmetric cell division (sporulation) of *B. subtilis* creates a smaller daughter cell (the forespore) that initially contains only 30 % of its copy of the chromosome;

the remaining 70 % must be translocated from the larger daughter cell by SpoIIIE, a hexameric, membrane-anchored DNA translocase [247, 248]. It was originally thought, based on similarities with bacterial conjugation systems, that SpoIIIE forms a single aqueous channel between the mother cell and forespore [247]. Later studies suggested, on the contrary, that the septum is closed and two SpoIIIE hexamers jointly form a channel across both membranes through which the DNA passes, based on inability of GFP expressed specifically in the mother cell to diffuse to the forespore [248]. Since both models predict different SpoIIIE copy numbers at the translocation septum, they could be resolved by accurately counting of SpoIIIE monomers.

Finally, at the cellular level, proteins and protein complexes can form higher order structures and super-resolution can provide deeper insight into the biological effect of such structures from the spatiotemporal ordering of its constituent proteins. For instance, *E. coli*'s chemotactic clusters—which allow the sensing of gradients of small molecules—involve tens of thousands of receptor proteins [249]. These clusters are positioned in an apparently periodic fashion on the membrane [138]. It had been suggested, from time-lapse fluorescence microscopy, that receptor proteins are in fact inserted at random in the cell membrane but later migrate to pre-existing anchor sites [138]. Other models proposed that this periodicity arises spontaneously from the stochastic nucleation and merging of clusters [138]. Greenfield et al. suggested that studying the protein number distribution per cluster could offer insights in the mechanism by which they are formed [138]. We will revisit the type of insight afforded by super-resolution to this question later.

III.3.1 Counting from fluorescence intensity

Proteins localized in small clusters can be counted without the need to spatially resolve them. To do so, we may estimate the number of photons collected and divide through by the mean number of photons emitted by each fluorophore. This mean number (the fluorophore's photon budget) depends not only on the excitation used but also, more crucially, on specific cellular conditions, in addition to other properties such as possible fluorophore interactions.

As an example of this early approach, if clusters contain few fluorophores, a histogram of the cluster brightnesses may exhibit discrete peaks at multiples of a base value [250]. In such a case, this base value likely corresponds to the intensity of a single labeled protein and peaks observed at two, three, or more times this intensity correspond to clusters of two, three, or more proteins [250].

In a different approach, a calibration curve relating fluorescence intensity to fluorophore number is constructed by engineering arrays of, say, 12, 24, and 36 fluorophore binding sites and measuring the fluorescence intensity for each number of bound markers [251]. Crucially, such a method can be used to establish the existence of a nonlinear relationship between

fluorophore count and fluorescence intensity, that can be caused, for example, by interactions among fluorophores [251].

The precision of methods relying on total observed intensity is relatively low and relying on a standard mean fluorophore brightness is not without risks. For example, in 2006, Joglekar et al. GFP-labeled a number of yeast kinetochore proteins (where the kinetochore is the structure that links centromeric DNA to spindle microtubules) [252]. They relied on the fluorescence of a single protein within the complex, Cse4, as a GFP fluorescence standard as that protein was thought to exist in a single copy per complex [252]. However, later studies demonstrated that this assumption was incorrect: Cse4 may be present in 4 to 8 copies per centromere and the reported counts of all other proteins were thus underestimated by the same ratio [253, 254]. Such a difference disqualified earlier arguments indicating that Cse4 may be present in too small a quantity to maintain the necessary attachment points [253, 254].

Further biochemical studies (protection assays) suggested that one of these proteins (centromere protein A, CENP-A) was, in fact, present at the levels suggested in the Joglekar study [255]. This time, it was argued that the larger numbers observed by the Coffman and Lawrimore studies [253, 254] arose from the inclusion in their counts of "unincorporated" labeled CENP-As, i.e., those not part of the structure itself but simply lingering in the structure's vicinity, possibly due to lower incorporation efficiency of labeled CENP-A. Since biochemical studies are not devoid of artifacts either, this controversy remains open to this day [256], and should serve as a reminder that the biological question is not to know how many fluorescent proteins are present somewhere but how many of the underlying proteins are actually participating in the process of interest.

III.3.2 Counting by photobleaching using diffraction limited data

An alternative approach is to rely on the stochastic photobleaching of single fluorophores [245]. More precisely, we rely on the observation that the times at which multiple active fluorescent proteins appearing within the same diffraction spot eventually photobleach are stochastic and thus likely different from one another. Thus, a time series of the total fluorescence signal will exhibit a stepwise decrease [245] with possible double-sized steps if two fluorophores simultaneously photobleach within the time scale of data acquisition.

So long as a majority of steps are resolvable, most steps should coincide with the photobleaching of a single fluorophore (or reversible transitions to and from dark states for blinking fluorophores). This is especially true toward the end of the photobleaching trace where the odds of two simultaneous photobleaching events are comparatively low. Thus,

given an estimate for the single fluorophore intensity drop, the number of labeled proteins present at the start of the trace can be estimated as the ratio between the initial intensity and the fluorescence drop arising from a single photobleaching event.

Leake et al. applied this method to study MotB, a component of the stator of *E. coli*'s flagellar motor, concluding that 22 ± 6 copies were present per complex [245]. However, this method also suffers from low precision as the noise level at the start of the trace is high and the initial intensity is therefore poorly defined.

Instead of trying to resolve fluorescence decrease steps, which may be challenging, one may compare the evolution over time of the total intensity of a collection of fluorophore spots (which is decreased by any photobleaching event) to the evolution of the number of spots, within that same collection, which have not completely photobleached yet (which decreases only when all the fluorophores within a given spot have photobleached). The slower the decrease of the number of spots relative to the decrease of the intensity, the larger the number of fluorophores per spot [257]. Yet another approach is to count molecules by means of photon arrival statistics [258, 259]. This technique exploits the photon antibunching effect, which essentially states that a single emitting quantum system (a fluorophore in this case) emits photons one at a time. Therefore, if the temporal resolution of the detector is sufficiently fine, photons detected at the same time can only originate from different emitters. In its most recent implementation [259], photon counting statistics were gathered and then a nonlinear regression with a Levenberg-Marquardt algorithm was used to back out the number of emitters, i.e., molecules of interest. However, this method is limited to counting up to around 20 molecules, largely because of error introduced by blinking and photobleaching effects.

A more promising albeit more difficult approach is to attempt to identify and count all individual photobleaching steps, and use their number as an estimate of the protein count [186]. For example, Ulbrich et al. studied the composition of a membrane-bound receptor in *X. laevis* oocytes, that was known to form tetramers [186]. The number of steps in each photobleaching step was visually estimated. Interestingly, the distribution of the number of steps resolved (1 to 4) is well fitted by a binomial distribution, consistent with a model that only about 80 % of the labels are ever fluorescent.

In the sections that follow we will explore theoretical approaches that have been proposed to locate photobleaching steps that can be resolved.

III.3.2.1 Hidden Markov modeling of photobleaching time series

Even before Ulbrich's original experiments, Messina et al. [260] proposed to determine the number, N, of fluorophores using HMMs [261, 44], where each state coincides with a combination of states for each individual fluorophore.

The large number of states in this model is suitably shrunk by exploiting the fact that states with the same number of bright fluorophores are indistinguishable, leading to a formulation where each state corresponds to a number n of active fluorophores. For instance, the transition from a state with n active fluorophores to $n-1$ has a rate equal to n times the transition rate to the dark state (as any of the n fluorophores could go dark) and, similarly, the transition from the state with n active fluorophores to $n+1$ has a rate equal to $N-n$ times the recovery rate from the dark state (as any of those $N-n$ fluorophores could recover) [260].

Standard maximum likelihood techniques [183] were then applied to compute the likelihood corresponding to each total number of fluorophores N. As there is no penalization for overfitting, this likelihood can only increase for increasing N; however, it is expected to plateau once the true number of underlying fluorophores is reached.

This method's original implementation was applied to time-correlated single photon counting experiments; that is, a setup where stochastic arrival times of each individual photon is measured [260], rather than the more common setup where an average intensity is measured by integration over a longer period. In such a case, the source of noise arises from the existence of "background" photons not associated with a fluorophore of interest [262], as well as from the stochasticity of the arrival times of the "true" photons. However, the approach of Messina et al. can also be directly adapted to the case where an average intensity is measured [263]. The authors suggest that up to 30 fluorescent dyes may be counted using such a technique [260].

III.3.2.2 Step-finding algorithms in counting by photobleaching

Without characterizing the kinetics of photobleaching, it is also possible to rely on classical step-finding algorithms to count the number of photobleaching events in a time trace [264–266]. The problem of locating sharp discontinuous changes in noisy data, the purview of step-finding algorithms, is a general problem across science that has been investigated across single molecule biophysics (see part I for another application of step-finding algorithms).

III.3.2.2.1 Edge-preserving smoothing

Many step-finding algorithms start from an initial filtering or downsampling of the data [267]. Although linear filters, where each data point is replaced by a weighted average of the neighboring data points within a specified window, are easily implemented, they also tend to blur or smooth out true transitions in the data. In particular, multiple temporally close transitions may become "merged" into a single transition [268].

To avoid this effect, Chung and Kennedy [268] proposed (for the purpose of resolving state transitions in patch-clamp experiments) a nonlinear filter, whereby the weight given to a

neighboring point during the filtering depends on how well it predicts the current observation, an approach known in the image-processing field as edge-preserving smoothing.

More precisely, for each data point $y(t)$ in a trace, we consider $2K$ "predictors" of "order" $-K, -(K-1), \ldots, -1, +1, \ldots, K-1, K$, namely the averages of the i ($1 \leq i \leq K$) previous or i future data points:

$$\langle y \rangle_{-,i}(t) = \frac{y(t-\delta t) + \cdots + y(t-i\delta t)}{i} \quad \text{and} \quad \langle y \rangle_{+,i}(t) = \frac{y(t+\delta t) + \cdots + y(t+i\delta t)}{i}. \tag{III.30}$$

The squared error provides a metric quantifying the predictor's quality:

$$\Delta_{-,i}(t) = \left(\langle y \rangle_{-,i}(t) - y(t) \right)^2 \quad \text{and} \quad \Delta_{+,i}(t) = \left(\langle y \rangle_{+,i}(t) - y(t) \right)^2. \tag{III.31}$$

The Chung-Kennedy filter is then computed by weighting the predictor of order i by the inverse pth power of the average badness of the predictors of the same order i but considered over the M preceding (for negative orders) or following (for positive orders) data points,

$$CK(t) = \frac{1}{Z} \sum_{i=1}^{K} \left[\left(\sum_{j=0}^{M-1} \Delta_{-,i}(t-j\delta t) \right)^{-p} \langle y \rangle_{-,i}(t) + \left(\sum_{j=0}^{M-1} \Delta_{+,i}(t+j\delta t) \right)^{-p} \langle y \rangle_{-,i}(t) \right] \tag{III.32}$$

(where the denominator is simply a normalization factor).

This filter possesses three parameters (K, M, and p), which are tuned empirically. The authors demonstrate that an appropriate choice of the parameters leads to a reduction of the noise without distorting sharp transitions [268]. More quantitatively, the effect of various filters on the quality of various step-finding algorithms has been the subject of a comparative study by Carter et al. [267] finding that a properly (manually) tuned Chung-Kennedy filter exhibited better performance than mean or median filtering.

III.3.2.2.2 Segmenting the trace

Regardless of whether (and how) the data is smoothed to facilitate step-finding, the essential part of step-finding is to segment a trace into "approximately constant" regions separated by a step. Two approaches are possible: "bottom-up", where small regions are merged together on the basis of value closeness, and "top-down", where the whole trace is progressively split into separate regions.

An example of the bottom-up approach was proposed by McGuire et al. [264]. Briefly, starting from the beginning of a trace, data points are progressively added to a running window until the value of the fluorescence moves outside of a small range centered at the current window mean. When this occurs, the current window is terminated and a new

window started. After running this process on the whole trace, it is repeated on the resulting "leveled" trace until the levels have converged. This whole process is then iterated (starting from the "leveled" trace) using progressively wider window ranges [264].

Conversely, an example of a "top-down" approach is provided by a mathematical idealization of the white noise assumption: the goal is to find the piecewise constant signal $\hat{y}(t)$, containing N discontinuities (steps), that minimizes the mean square error, $\left\langle (y(t) - \hat{y}(t))^2 \right\rangle$. Because it is computationally intractable to test all possible combinations of step numbers and their coinciding locations, the number of locations scaling as the number of time points raised to a power equal to the number of change points, Kalafut and Visscher [83] proposed to iteratively add change points one after another, each of them at the position that decreases the mean square error the most. At each iteration, one only needs to check, for each time point, the decrease in mean square error if the next change point was inserted there. Even a naive implementation of this approach only exhibits a complexity proportional to the product of the number of time points by the number of change points. We note however that recent theoretical developments have provided efficient exact algorithms to solve this problem through a careful pruning of the solution tree [269, 270].

As a fit's mean square error can only decrease as more steps are added, "top-down" approaches additionally require an explicit penalty against overfitting. Kalafut and Visscher [83] propose the use of the Bayesian information criterion (BIC) [271]. However, as acknowledged by the authors, this penalty is typically insufficient and leads to overfitting of the data.

For particular applications, it is always possible to create better step-finding algorithms directly informed by the physics that dictates the noise properties of the problem.

For example, a Bayesian algorithm specifically applied to counting by photobleaching is presented by Tsekouras et al. [266]. In this method, priors and likelihoods are specifically informed by the physics dictating that noise properties should vary stochastically, on the basis of the number of active fluorophores, and that the number of overlapping blinking and photobleaching events have different a priori expectations based on the length of the time trace and the stochastic nature of the photobleaching process.

With this information at hand, Tsekouras et al. arrive at a "top-down" method, more precisely, a marginal posterior for the entire trace, that can be used as a criterion to locate photobleaching steps. The method succeeds in avoiding the overfitting problem arising from assumptions of constant noise across a dataset and, according to the authors, scores correctly dozens or even hundreds of steps, provided enough data points are present between successive steps to avoid small number statistics problems.

III.3.3 Counting by blinking correction

As suggested earlier, super-resolution microscopy, and PALM in particular, are seemingly well suited for counting, as the molecular photoactivation times are, by design, as temporally separated as possible. Thus, the number of fluorescence events or bursts detected within a diffraction-limited spot should, in theory, match the number of active fluorophores within that spot. This is only in principle true if fluorophores do not blink.

However, even in the presence of blinking, counting remains possible so long as consecutive bursts originating from the blinking of a single fluorophore can be grouped together. Indeed, threshold methods—described in greater detail below—were used, for example, to study the size distribution of *E. coli*'s chemotactic cluster [138].

III.3.3.1 Threshold methods for counting from PALM data

In the very first implementation of PALM, Betzig et al. [57] acknowledged the need for such a grouping and applied a purely empirical threshold ("blinking correction time") to merge events appearing within the neighboring pixels and separated by no more than three dark frames.

Annibale et al. [272] further studied the blinking kinetics of the widely used mEos2 photoactivatable fluorescent protein (PA-FP), imaged on a coverslip in vitro. By lowering the fluorophore density, the authors could ensure that each fluorescent event corresponded indeed to a single fluorophore (further confirmed by the absence of multi-step photobleaching) [272]. The authors found that roughly half of the molecules reactivated after entering a dark state, i.e., blinked. Recovery times from the dark state were found to be multiexponentially distributed (similar to observations on other PA-FPs [273]) and regularly lasted as long as tens of seconds. This observation thus raised the concern that earlier studies may have misinterpreted large number of blinking events as evidence for protein oligomerization [272].

The authors thus proposed two methods to correct for this blinking. Either the blinking correction time could be empirically increased or, perhaps more interestingly, the authors found that recovery from the dark state could be accelerated, and thus the blinking correction time kept low, through continuous illumination by the photoactivation laser. Thus, they recommended the use of a continuous photoactivation scheme, rather than a pulsed activation scheme where the photoactivation laser is alternatively turned on for a brief period of time, then kept off while the fluorescence of the activated subset is collected [272].

In order to maximize the accuracy of the estimated number of events, Lee et al. offered an alternative selection strategy for selecting the blinking correction time [60]. First, the authors suggested a scheme by which the photoactivation laser power is tuned in order to ensure a near-constant number of photoactivation events per unit time. Such a strategy ensures

maximal separation between active active fluorophores in times and thus minimizes the probability that two fluorophores be simultaneously active.

In short, the scheme was devised by first considering the total number of molecules, $N(t)$, yet to be photoactivated at time t with instantaneous activation rate $k(t)$ which was found to be proportional to the photoactivation laser power, $P(t)$. The number of molecules that photoactivate between times t and $t + \mathrm{d}t$ is $\mathrm{d}N = k(t)N(t)\mathrm{d}t$; thus, the number of molecules that photoactivate per frame could be kept constant by solving

$$\frac{\mathrm{d}N}{\mathrm{d}t} = k(t)N(t) \tag{III.33}$$

and selecting a $k(t)$ (or equivalently $P(t)$) that enforces a constant $\mathrm{d}N/\mathrm{d}t$ [60]. The authors subsequently generated a simulated time series corresponding to the blinking behavior of various fluorophore numbers. The number of events in each time series was then counted by using various possible values for the blinking correction time; the blinking correction time that achieved, for each given underlying number of fluorophores, the minimum mean bias was then tabulated. As could be expected, the more fluorophores in a single spot, the smaller the correction time that achieved unbiased counting. Since the correct number of fluorophores is initially unknown, the authors proposed to start from an arbitrary count, pick the corresponding correction time, count using that correction time and iterate. Overall, this approach was shown to exhibit an error of a few percent when counting up to a hundred molecules [60].

While this method introduces a blinking correction time (i.e., a threshold), it does not attempt to produce a "grouping" of events correctly identifying whether two events truly arose from the same molecule. In general, such a grouping may be impossible to attain, as the blinks arising from multiple molecules may be interlaced, or even overlap each other. In such a case, "undercounting" (the incorrect merging of events corresponding to two different fluorophores) is unavoidable. Instead, the blinking correction time was chosen (by using the simulations described above) so that these undercounts are exactly compensated by "overcounts" which are cases where a single fluorophore took a time longer than the blinking correction time to recover from its dark state [60].

III.3.4 Limitations of counting

Biological constraints complicate the counting problem. For instance, even if all proteins are labeled and are expressed in their native amount (which has been made possible by the advent of widespread genome editing systems), not all fluorophores mature [186, 274], nor will all photoconvertible fluorescent proteins successfully photoconvert [275]. Fundamentally, no algorithm can count proteins that never appear.

Various approaches have been proposed to quantify the percentage of proteins that properly activate. For example, by expressing a labeled human glycine receptor GlyR, whose known stoichiometry of three α and two β subunits could be used as a reference, in *X. laevis* oocytes and counting them either by stepwise photobleaching or by blinking correction, Durisic et al. found that across a wide range of photoconvertible fluorescent proteins, only 40 % to 80 % of the proteins successfully photoconverted [275]. Likewise, Wang et al. expressed a dozen different fluorescent proteins in *E. coli* and compared the number of events collected in a PALM experiment, corrected through division by the mean number of blinks per molecule, to an estimate of the actual number of fluorescent proteins expressed, obtained by quantitative Western blotting [180]. They found an even lower detection efficiency for fluorescent proteins: only 1 % to 20 % of them successfully photoconverted. Such limitations need to be taken into account while comparing the accuracy of counting methods: minor gains in the theoretical accuracy of counting will only matter if the global accuracy of the count is not limited by experimental considerations.

Chapter III.4

Conclusion

More than ten years after the publication of the seminal papers on super-resolution [56–58], single molecule super-resolution microscopy has become a standard lab technique now widely available across major imaging facilities. It has been used to study a number of biological targets just below the resolution of diffraction-limited microscopy, such as microtubules [276], mitochondria [277], and the nucleopore complex [278, 279]. Theoretical developments in interpreting super-resolution experiments, and single molecule experiments more broadly, have ushered data-driven methods into the physics and chemistry mainstream [280]. While studies in live cells are motivating more general theoretical approaches borrowing heavily from statistical advances [281] now feasible due to computing power, quantitative analysis efforts have also helped identify clear challenges standing in the way of greater modeling accuracy. Novel experimental methods have begun addressing some of these challenges such as phototoxicity [282] and image distortions in thick heterogeneous samples [283] though other key challenges such as labeling density [284] and environment-dependent photophysical properties [285] remain.

As this is a review of analysis methods, we highlight three broad directions that have been the focus of recent theoretical efforts. The first is on joint methods that simultaneously, and thus self-consistently, solve many problems at once such as problems in interpretation and counting [150]. Such efforts reduce the number of user-dependent postprocessing steps albeit at a heavier computational cost. The second introduces problem-specific models [286], priors (whether theoretically [266] or experimentally [276] motivated), and algorithms [287, 288] suited to the particularities of the physical (or photophysical) challenge to reduce the computational burden and improve the prediction accuracy. The third is focused on generalizing models to accommodate the data's complexity [289, 290].

The picture of life emerging from breakthrough experimental techniques and analysis methods is one far richer in structural features, dynamics, and stochasticity than we could have conceived of even a decade ago. We envision a future in imaging where a combination of experiments and principled analysis provide a compelling narrative into the chaotic journey of life from the level of single molecules upward.

Part IV

Super-resolution imaging of protein-protein interactions through bimolecular complementation and photoactivated localization microscopy

This work was done in collaboration with Alyssa Rosenbloom, Sang-Hyuk Lee, Jae-Yen Shin, and Carlos Bustamante.

Super-resolution microscopy, a powerful technique for determining protein localization, cannot discriminate between casual protein-protein proximity and stable interactions. Here, by developing split-photoconvertible fluorescent proteins that recover photoconversion and fluorescence properties upon complementation, we combine the absolute localization ($\sim$ 20 nm) of photoactivated localization microscopy (PALM) with the relative localization ($<$ 10 nm) of bimolecular fluorescence complementation (BiFC) for in vivo super-resolution visualization of stable protein-protein interactions between two subunits of the mammalian ATP synthase.

Chapter IV.1

Introduction

A common method for determining protein-protein interaction in vivo is co-localization using two-color fluorescence microscopy. Classical microscopy is limited by diffraction but super-resolution methods, such as photoactivated localization microscopy (PALM), circumvent this limit by combining temporal and spatial separation of individual fluorescent events to determine absolute localization of single molecules with a resolution of $\sim 20\,\text{nm}$ [56–58]. Several two-color super-resolution methods are described in the literature, but none can differentiate between transient proximity, limited by overlapping localization uncertainties, and stable protein-protein interaction [179, 240].

Bimolecular fluorescence complementation (BiFC) makes use of certain fluorescent proteins (FPs) that, when divided into two non-fluorescent fragments, can re-associate into a fully fluorescent protein complex [291]. Successful complementation requires two fragments be held in close proximity ($< 10\,\text{nm}$) for an extended period of time, generally by fusion via short linkers to two stably interacting proteins. The formation of the fluorescent complex, driven by these interacting proteins, competes kinetically with the irreversible misfolding of the individual fragments, which prevents the apparition of fluorescence [292]. The amount of fluorescence observed correlates with the strength of the interaction, down to $K_D \sim 1\,\text{mM}$ [293].

Standard split-FPs have been used previously in conjunction with stimulated emission depletion microscopy (STED) [294]. PALM, while offering superior resolution, requires photoactivatable or photoswitchable fluorescent proteins (PA- or PS-FPs). Thus, we hypothesized that stable protein-protein interactions could be determined at super-resolution by combining the relative resolution of BiFC with the absolute resolution of PALM through the use of split PA- or PS-FPs. In this scheme, each individual fluorescent event ($\sim 20\,\text{nm}$) would then represent two proteins that stably interact and co-localize within $< 10\,\text{nm}$ of each other. So far, Dronpa is the only photoconvertible fluorescent protein for which successful splitting has been reported [295]. Here we show that PAmCherry1, Dendra2, and a Dronpa

variant, rsKame, can be split and, upon complementation, maintain their fluorescence and photoactivatablity [296–298].

We explored super-resolution microscopy with BiFC-PALM by labeling two subunits (OSCP and b) of the mammalian ATP synthase stator stalk with N- and C-terminal halves of PAmCherry1 respectively. Here, we demonstrate that the stable interaction between these two subunits, previously shown by FRET in *S. cerevisiae*, allows for the complementation of split-PAmCherry1, while random collisions between freely diffusing molecules are insufficient [299]. In addition, we were able to successfully obtain individual fluorescent events at $\sim$ 20 nm resolution, each representing a single pair of interacting subunits in an ATP synthase complex.

Chapter IV.2

Results

In order to assess the suitability of two-color PALM for studying protein-protein interaction at super-resolution, we chose a system previously characterized by FRET. The F_o-F_1 ATP synthase is a highly conserved oligomeric complex that uses the proton gradient across the mitochondrial inner membrane to synthesize ATP. Proton passage through the F_o pore drives the rotation of a central stalk through the F_1 domain, which is held in place by a peripheral stalk [300]. In *S. cerevisiae*, two of the subunits of the peripheral stalk, b and OSCP, have been shown to interact by FRET measurements between subunit b-GFP and OSCP-BFP fusions [299]. We co-expressed mammalian subunit b fused to rsKame and OSCP fused to PAmCherry1 on a pIRES vector in COS-7 cells for PALM imaging. For both fusions, a 12 amino acid linker (GSSGGGGSGGGG) was used. After 36 hours of expression at 37 °C, cells were fixed with 1 % formalin in 1× PHEM. We first imaged rsKame under 488 nm excitation (6 W/mm^2) and 405 nm photoactivation (0 mW/mm^2 to 10 mW/mm^2) for about 5 min and then PAmCherry1 under 561 nm excitation (22 W/mm^2) and 405 nm photoactivation (0 W/mm^2 to 3 W/mm^2) for about 8 min [277]. 100 nm gold beads were used as fiducial markers for drift correction. PALM movies were analyzed with a custom MATLAB-based software package [60]. Individual fluorescent events were identified and localized at $\sim$ 20 nm resolution, both for rsKame and for PAmCherry1 (figure IV.1a,b), showing co-localization of red and green fluorescence, and thus of subunit b and OSCP, as expected (figure IV.3a). However, when we targeted instead PAmCherry1 to the mitochondrial matrix with the cleavable mitochondrial matrix localization sequence (MLS) of CoxVIII instead [301], while keeping the subunit b-rsKame fusion, and imaged COS-7 cells under the same conditions, co-localization of red and green fluorescence was also observed, despite the lack of interaction between the fusion partners of the two FPs (figure IV.3b). Thus, protein-protein interaction cannot be proven by fluorescence co-localization alone, even at super-resolution. This led us to develop BiFC-PALM for such applications (figure IV.3c).

The splitting of various FPs such as EGFP and mCherry1 between β-sheets 7 and 8 has been characterized previously [291, 302]. As PAmCherry1, mEos2, Dendra2 and Dronpa variant

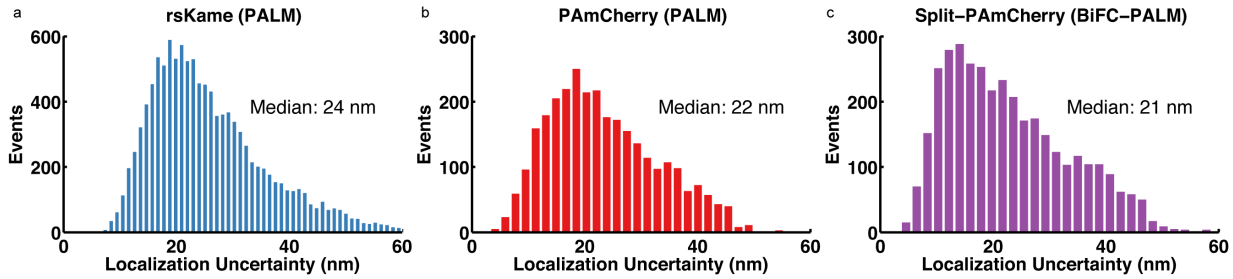


Figure IV.1:
Localization uncertainty of (a) full-length rsKame (PALM image shown in Fig 1a), (b) full-length PAmCherry1 (PALM image as shown in Fig 1a) and (c) complemented split-PAmCherry1 (BiFC-PALM image as shown in figure IV.4c), computed for all the molecules in each field of view using the theoretical estimation [214].

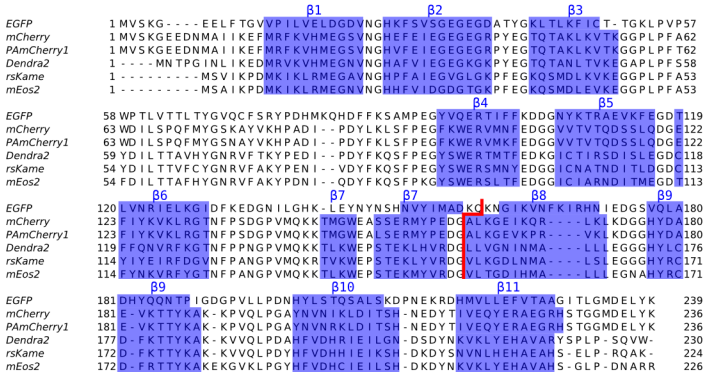


Figure IV.2:
Sequence alignment between fluorescent proteins EGFP, mCherry, PAmCherry1, Dendra2, rsKame, and mEos2 demonstrates a conserved splitting site (indicated in red) between β7 and β8. Blue boxes indicate β-sheets.

rsKame share similar β-barrel structures to EGFP and mCherry1, we also split them between β7 and β8 (figure IV.2) [277, 296–298, 300]. In order to test whether complementation of the PA- or PS-FPs restores the photoconversion and the fluorescence properties of the full proteins, we used the self-dimerization of either EGFP or TagBFP under over-expression conditions [302]. N- or C-terminal halves of the split-PA- or PS-FPs were fused to EGFP or TagBFP, using the 12 amino acid linker described above, and co-transfected into EpH4 cells. Cells were allowed to express proteins at 37 °C for 24–48 hours after which the incubation temperature was lowered to 28 °C for 12 hours, to allow for complementation. For cells expressing the split-PAmCherry1-, split-mEos2-, or split-Dendra2-EGFP fusions, EGFP expression was confirmed by 488 nm illumination. Under 561 nm illumination, red fluorescence was observed in cells expressing split-PAmCherry1 or split-Dendra2 only after photoactivation by 405 nm for $\sim$ 10 s (arc lamp), indicating successful complementation (figure IV.3d). No fluorescence could be recovered in cells expressing split-mEos2. Likewise, for cells expressing split-rsKame-TagBFP fusions, TagBFP expression was confirmed by 405 nm illumination. Green fluorescence from rsKame was switched off by 488 nm illumination and recovered after photoactivation by 405 nm illumination (figure IV.3d). As we were able to recover fluorescence in BiFC-PA-FPs or BiFC-PS-FPs that are either irreversibly photoactivatable by UV-induced peptide cleavage (PAmCherry1 and Dendra2) or reversibly photoswitchable through a cis to trans isomerization of the chromophore (rsKame), functional complementation does not interfere with either photoactivation mechanism. We do not know why mEos2 did not functionally complement.

Having shown that photoactivatable fluorescence can be recovered upon complementation of split-PAmCherry1, we then labeled OSCP and subunit b with the N- and C-terminal halves of split-PAmCherry1 respectively, using the 12 amino acid linker described above, and co-expressed them on a pIRES vector (figure IV.4a). In vivo complementation of split-PAmCherry1 was confirmed by imaging transfected COS-7 cells after 36 hours of expression at 37 °C and 12 hours at 28 °C, to allow for complementation. Mitochondrial networks were pre-stained with MitoTracker Green and imaged by illumination with 488 nm. After 90 s of 405 nm photoactivation (7.3 W/mm^2) we observed mitochondrially localized red fluorescence from complemented PAmCherry1 under 561 nm excitation (figure IV.4b), both in live cells and in cells fixed with 1 % formalin in 1× PHEM.

Identically prepared COS-7 cells were also used for BiFC-PALM imaging. Split-PA-mCherry1 was imaged under 561 nm excitation (22 W/mm^2) and 405 nm photoactivation (0 W/mm^2 to 3 W/mm^2) until complete photobleaching. Again, individual fluorescent events were identified and localized at $\sim$ 20 nm resolution (figure IV.4c, figure IV.1c), yielding super-resolution images of a quality similar to those obtained by keeping only the red channel of the previous two-color PALM experiments (figure IV.4d). However, this time, each event represented a single ATP synthase complex containing both labeled subunit b and OSCP in close proximity (< 10 nm).

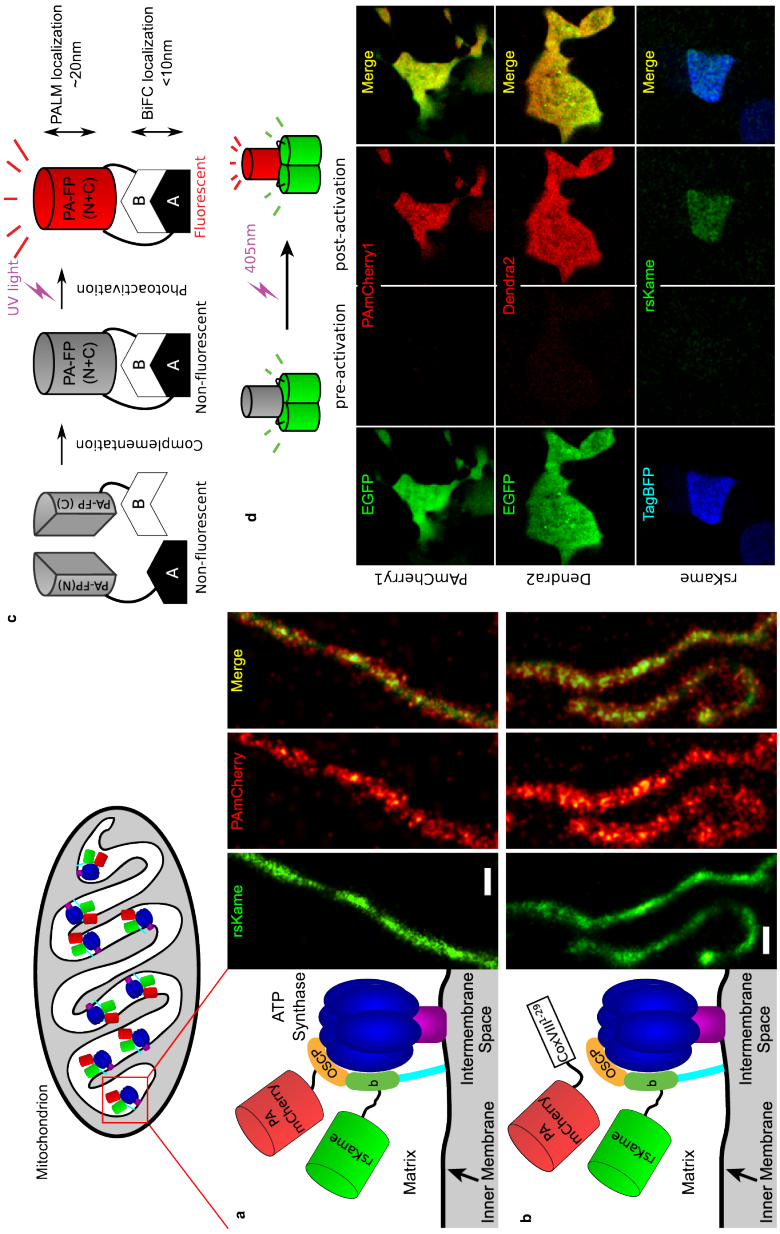


Figure IV.3: (See next page.)

(continued)

(a) Interacting proteins OSCP and subunit b, respectively labeled with PAmCherry1 (red) and rsKame (green), co-localize when imaged by two-color PALM (merged image).

(b) However, super-resolution co-localization is also observed when PAmCherry1 is targeted to the mitochondrial matrix instead using the MLS of CoxVIII (red) while subunit b is still labeled with rsKame (green), despite the lack of protein-protein interaction (merged image).

(c) A stable interaction between proteins A and B, respectively fused to the N- and C-terminal halves of a split PA-FP, can drive the complementation of the PA-FP. Photoactivation of the fully complemented PA-FP results in fluorescence upon excitation.

(d) Stable EGFP or TagBFP dimerization at high expression levels leads to the complementation of split PAmCherry1, Dendra2 and rsKame in EpH4 cells. EGFP was fused to both halves of PAmCherry1 and Dendra2, and TagBFP to both halves of rsKame. Expression of EGFP and TagBFP were detected by excitation at 488 nm and 405 nm, respectively. Prior to photoactivation by 405 nm illumination, no fluorescence was detected from complemented PAmCherry1, Dendra2 and rsKame. After a 30s photoactivation by 405 nm (arc lamp), fluorescence was detected by 561 nm or 488 nm excitation.

Scale bars, 0.5 µm (a, b).

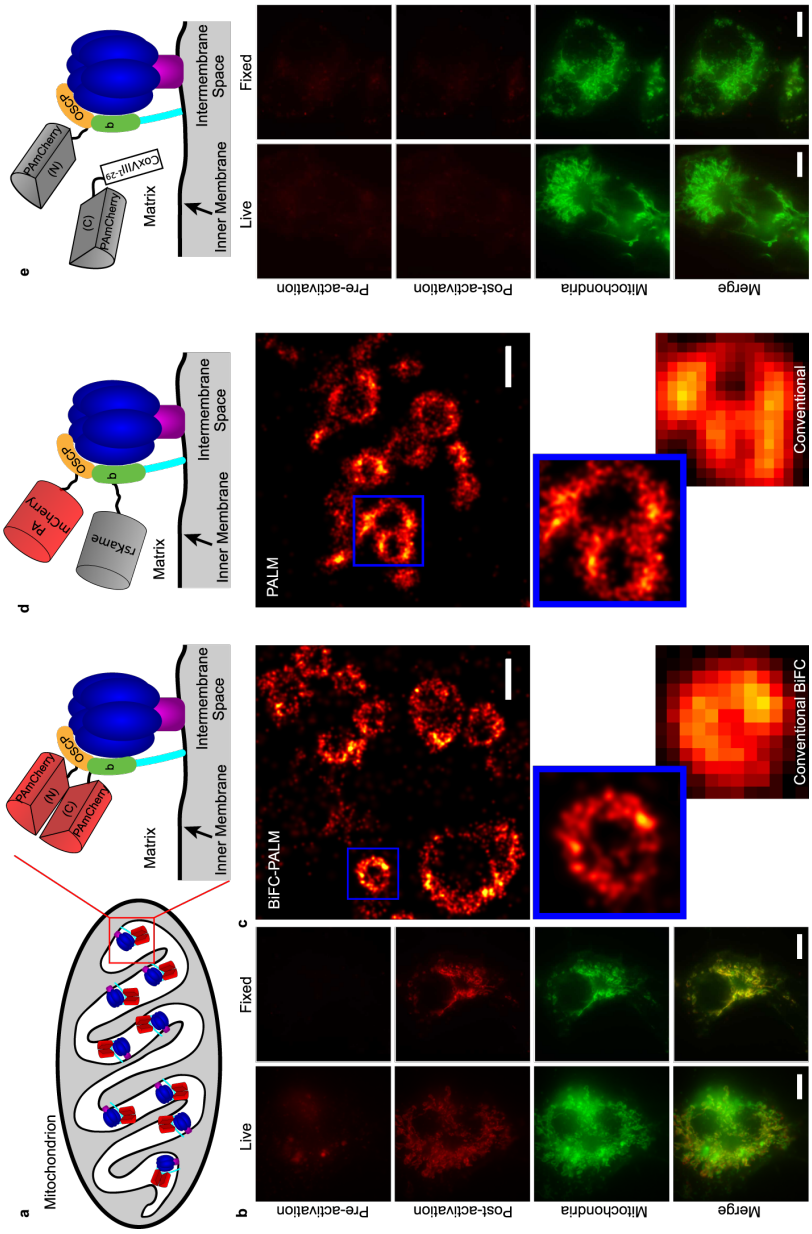


Figure IV.4: (See next page.)

(continued)

(a) The stable interaction between subunits OSCP and b of the ATP-synthase enables their fusion partners, respectively the N- and C-terminal halves of PAmCherry1, to complement into a photoactivatable complex.

(b) Photoactivatable fluorescence of complemented PAmCherry1 is detected by epi-fluorescence in live and fixed COS-7 cells, and is localized to the mitochondria. Under 561 nm illumination, no fluorescence is detected prior to activation, but fluorescence was detected post activation by 405 nm illumination for 90 s. The mitochondrial networks were pre-stained with MitoTracker Green.

(c) Super-resolution imaging using complemented PAmCherry1 in fixed COS-7 cells (top and blue box) offers a dramatic resolution improvement over conventional epifluorescence imaging (bottom).

(d) As a reference, a PALM image was reconstructed using only the signal from the red channel (OSCP-PAmCherry1) of the previous two-color PALM experiments (top and blue box). The number of events detected and the resolution improvement over diffraction-limited epifluorescence (bottom) are similar to those obtained by BiFC-PALM.

(e) Random, transient collisions between the split halves of PAmCherry1 are not sufficient for complementation: when PAmCherry1(C) was fused to the MLS of CoxVIII instead and co-transfected with PAmCherry1(N) fused to OSCP in COS-7 cells, no fluorescence is detected in live or fixed cells, before or after activation, under the same imaging conditions as in (b).

Scale bars, 10 µm (b,e), 0.5 µm (c,d).

We confirmed that complementation specifically requires long-term stable proximity by targeting the C-terminal half of PAmCherry1 to the mitochondrial matrix, again using the MLS of CoxVIII, while keeping the N-terminal half fused to OSCP (figure IV.4e). Under the same expression and imaging conditions as above, no red fluorescence was observed even after photoactivation, demonstrating that stable interactions, rather than random collisions, are necessary for successful complementation.

Chapter IV.3

Discussion

We have demonstrated that photoactivatable (PAmCherry1, Dendra2) and photoswitchable (rsKame) fluorescent proteins can be split and recover fluorescence and photoconversion properties upon complementation. We also demonstrated that complementation of split-PAmCherry1 requires a stable interaction of their respective fusion proteins, such as subunits b and OSCP of ATP synthase. In the absence of interacting fusion partners, the split halves of PAmCherry1 cannot complement, even if present at a significant concentration in the small volume of the mitochondrial matrix, most likely because the transient collisions of the split halves are too short and random. Once complemented, split-PAmCherry1 can be successfully used for super-resolution fluorescence imaging of protein-protein interactions. Specifically, it was possible to observe in situ the relative localization of subunit b to OCSP within $< 10\,\text{nm}$ by complementation of split-PAmCherry1 while obtaining the absolute localization of each interacting pair within $\sim 20\,\text{nm}$.

BiFC directly probes protein-protein interactions up to a range of $\sim 10\,\text{nm}$ and a stability of $K_D \sim 1\,\text{mM}$, but, like conventional FRET, cannot determine absolute molecular positions better than the diffraction limit. Meanwhile, super-resolution microscopy can only infer protein-protein interaction through colocalization, but a single PALM voxel of $\sim 20\,\text{nm} \times 20\,\text{nm} \times 20\,\text{nm}$ is large enough to contain hundreds of proteins, that would all appear co-localized without necessarily interacting. By contrast, BiFC-PALM combines both the relative localization of BiFC and the absolute localization of PALM. This approach permits definitive determination of protein-protein interaction within $< 10\,\text{nm}$ of each other as well as the absolute localization of the proteins within $\sim 20\,\text{nm}$ inside the cell.

Chapter IV.4

Materials and methods

IV.4.1 Cloning

pPAmCherry1^{N}-EGFP was built by ligating a KpnI/XhoI fragment, PAmCherry1$^{1\text{-}160}$, and a XhoI/ApaI fragment, EGFP into pcDNA-5/TO (Invitrogen). A 12 amino acid linker, GSSGGGGSGGGG, was added to the N-terminus of EGFP by PCR. pEGFP-PAmCherry1^{C} was built by cloning a KpnI/XhoI fragment, EGFP, and a XhoI/ApaI fragment, PAmCherry1$^{161\text{-}236}$, into pcDNA-5/TO. The same 12 amino acid linker was added to the N-terminus of PAmCherry1^{C} by PCR. The same method was used to build pDendra2^{N}-EGFP, pEGFP-Dendra2^{C}, pmEos2^{N}-EGFP, pEGFP-mEos2^{C}, prsKame^{N}-TagBFP and pTagBFP-rsKame^{C}.

pATP5F1-PAmCherry1^{C}-IRES-ATP5O-PAmCherry1^{N} was built as follows. A PstI/XhoI fragment, ATP5F1 (Open Biosystems) and a XhoI/BamHI fragment, PAmCherry1$^{161\text{-}236}$, carrying a hexahistidine tag at its C-terminus, were ligated together, and then cloned into pIRES2-EGFP (Clontech), yielding pATP5F1-PAmCherry1^{C}-IRES-EGFP. Then, a MscI/XhoI fragment, ATP5O (Open Biosystems), and a XhoI/NotI fragment, PAmCherry1$^{1\text{-}160}$, carrying a FLAG tag at its C-terminus, were ligated together, and cloned into the previous vector, replacing EGFP. Both fusion proteins included the same 12 amino acid linker, added by PCR to the N-terminal of each PA-mCherry1 fragment.

pMLS-PAmCherry1^{C}-IRES-ATP5O-PAmCherry1^{N} was built as follows. The mitochondrial localization sequence of subunit VIII of human cytochrome c oxidase, composed of its first 29 amino acids, and a 3 amino acid linker, GGG, were added to the N-terminus of PAmCherry1$^{161\text{-}236}$ by nested PCR, yielding a PstI/XhoI fragment, carrying a hexahistidine tag at its C-terminus. This fragment was cloned into pATP5F1-PAmCherry1^{C}-IRES-ATP5O-PAmCherry1^{N}, replacing ATP5F1-PAmCherry1^{C}.

pATP5O-PAmCherry1-IRES-ATP5F1-rsKame was built as follows. A PstI/XhoI fragment, ATP5O, and a XhoI/BamHI fragment, PAmCherry1, carrying a hexahistidine tag at its C-terminus, were ligated and cloned into pIRES2-EGFP, upstream of the IRES, yielding

pATP5O-PAmCherry1-IRES-EGFP. Then, a MscI/XhoI fragment, ATP5F1, and a XhoI/NotI fragment, rsKame, carrying a FLAG tag at its N-terminus, were ligated and cloned into the previous vector, replacing EGFP. Both fusions contained the same 12 amino acid linker, added by PCR to the N-terminal of each fluorescent protein.

pMLS-PAmCherry1-IRES-ATP5F1-rsKame was built by cloning a PstI/BamHI fragment, MLS-PAmCherry1, also built by nested PCR and carrying a hexahistidine tag at its C-terminus, into pATP5O-PAmCherry1-IRES-ATP5F1-rsKame, replacing ATP5O-PAmCherry1.

In order to promote translation, a Kozak consensus sequence, GACACC, was added to the N-terminus of each construct before the start codon. PAmCherry1 and Dronpa were generous gifts from the Liphardt lab (University of California, Berkeley). TagBFP was a generous gift from the Hariharan lab (University of California, Berkeley). mEos2 was a generous gift from Sean McKinney. rsKame was developed in the Bustamante Lab. Dendra2 was purchased from Clontech Laboratories, Inc.

IV.4.2 Cell culture, transfection and protein expression

Two different cell lines were used: mouse EpH4 (mammary gland epithelial) and monkey COS-7 (kidney). EpH4 cells, originally isolated by C. Roskelley, were cultured in Dulbecco's Modified Eagle's Medium (DMEM) with 2 % FBS, 5 mg/L insulin and 50 mg/L gentamicin. COS-7 cells, purchased from the University of California, Berkeley Tissue Culture Facility, were cultured in DMEM with 10 % FBS, 1 % non-essential amino acids and 1 % penicillin/streptomycin. Cell reagents were purchased from the University of California, San Francisco Cell Culture Facility. Cells were grown at 37 °C with 5 % CO_2. Cells were transfected using lipid-based commercial transfection reagents: Lipofectamine 2000 (Invitrogen) for EpH4 cells and Xfect (Clontech) for COS-7 cells. Cells were transfected for 4–6 hours and then incubated for 24–48 hours (EpH4 cells) or 36 hours (COS-7 cells).

IV.4.3 Diffraction-limited microscopy

Cells were imaged on the Nikon Eclipse TE2000-S scanning confocal microscope (Nikon Instruments Inc.). Cells were plated on MatTek imaging tissue culture plates (MatTek Corporation). Cells were imaged in vivo in 1× PHEM. Cells transfected with EGFP or TagBFP were illuminated with 488 nm or 405 nm respectively. rsKame was photoactivated with 405 nm prior to illumination with 488 nm. PAmCherry1, Dendra2, and mEos2 were photoactivated with 405 nm prior to illumination with 561 nm. In cells expressing multiple fluorophores, images were taken consecutively then analyzed and overlaid using ImageJ.

IV.4.4 Super-resolution fluorescence microscopy

Super-resolution microscopy was performed with a homemade fluorescence microscope built with an Olympus IX71 body and an UAPO150xO/TIRFM-SP NA 1.45 objective. Three lasers—561 nm (Sapphire 561 nm, 200 mW; Coherent, Inc.), 488 nm (Cyan 488, 50 mW; Spectra Physics), and 405 nm (Cube 405 nm, 50 mW; Coherent, Inc.)—were combined, magnified, and focused on the back focal plane of the objective to illuminate a sample area of 55 μm × 55 μm. Fluorescence images were acquired by a low-noise, light-sensitive electron multiplying charge coupled device (EMCCD) (DV887ECS-BV; Andor) at 20 Hz with either a green emission filter (ET525/50; Chroma) for rsKame or a red emission filter (FF01-588/21; Semrock) for PAmCherry1, mEos2 and Dendra2. Focus drift was stabilized by active feedback control of a piezo-stage (CRIFF; ASI) whereas in-plane sample drift was corrected by tracking 100 nm gold fiducials (Microspheres-Nanospheres). The final imaging system provided 512 × 512 pixels of entire field of view, 107 nm/pixel, and 2D Gaussian point spread function (PSF) with $\sigma \sim 1.2$ pixel.

Raw data was analyzed in three steps by using a custom-developed PALM analysis software written in MATLAB.

EMCCD images were digitally filtered using a Laplacian of Gaussian filter to enhance their contrast and remove low-frequency background. Local maxima of the filtered image for which the peak value was greater than 6 times the background noise and the summed raw EMCCD counts over a 7 × 7 box was greater than 2000 were selected for further fitting.

Raw EMCCD counts were converted to photon counts. Around each selected local maximum, a 11 × 11 box was cropped. Molecular localizations were obtained by fitting the central 9 × 9 pixels, using the remaining outer pixels to estimate the mean and standard deviation of the local background fluorescence Maxima appearing in consecutive EMCCD frames within 1 pixel of each other were interpreted as being due to a single fluorophore; thus, for such cases, the sum of the whole stack of corresponding 9 × 9 boxes was fitted.

Super-resolution images were reconstructed from the single molecule localizations thus obtained, by representing each single molecule by a 2D Gaussian with a standard deviation given by the theoretical uncertainty of the localization of the corresponding PSF [57, 214]. The summing of the 2D Gaussians was accelerated by GPU-based parallelization (CUDA, Nvidia).

Conclusion

During our Ph. D., we had the opportunity to explore two different single molecule techniques: optical trapping and super-resolution microscopy. In both cases, we aimed at improving the resolution or the accuracy of the technique.

In the case of optical tweezers, we presented in part I a series of algorithmic and instrumental improvements that allowed us to resolve single base-pair stepping by slowed RNA polymerase in a robust fashion. We exploited this technique to obtain the distribution of dwell times taken by the enzyme between consecutive steps, and measure how this distribution is affected by the transcription inhibitor pyrophosphate. Many other small molecules affect the dynamics of transcription elongation and are of therapeutic interest [46], and it is our hope that the methodology we developed may be used to study at the finest level the dynamical effects of such small molecules on the elongation cycle.

In part II, we presented a methodology, relying on transcription through specifically constructed (namely, repeating) templates as well as novel data analysis methods, that allows us to accurately measure sequence-dependent dynamics of RNA polymerase, down to time scales where pauses and pause-free transcription become indistinguishable from a purely kinetic point of view. This framework allowed us to obtain important information regarding the dynamics of pause entry, the initial steps of backtracking, the role of the factor GreB, and the various effects that RNA secondary structures can have on RNA polymerase. But the framework is also more widely applicable, and is already being used by colleagues to study other sequence-dependent processes, such as the transcription by the eukaryotic RNA polymerase (Pol II) through nucleosomes.

It is our strongly held opinion that further improvements to the resolution and accuracy of optical tweezers may be achieved through careful modeling of the data (which cannot be decoupled from a deep understanding of the biological system of interest) and through the application of novel statistical methods.

As part of our work on super-resolution microscopy, we first presented in part III a review of current analysis techniques, both for achieving the highest possible localization accuracy, as well as for extracting quantitative information regarding protein counts from super-resolution data. Finally, we presented in part IV our work in engineering a split photoactivatable fluorescent protein that may be used for probing protein-protein interaction at high resolution while maintaining the localization accuracy advantages offered by super-resolution microscopy. Although this project was prematurely interrupted following the publication of similar findings by other groups [303, 304], we still hope that the efforts to develop general methods for probing protein-protein interactions (either with bimolecular fluorescence complementation, or with more flexible techniques such as Förster resonance energy transfer) with high localization accuracy will continue.

Bibliography

1. Crick, F. Central dogma of molecular biology. *Nature* **227,** 561 (1970).
2. Gruber, TM & Gross, CA. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.* **57,** 441–466 (2003).
3. Murakami, KS. Structural biology of bacterial RNA polymerase. *Biomolecules* **5,** 848–864 (2015).
4. Murakami, KS & Darst, SA. Bacterial RNA polymerases: the wholo story. *Curr. Opin. Struct. Biol.* **13,** 31–39 (2003).
5. Johnston, DE & McClure, WR. *Abortive initiation of in vitro RNA synthesis on bacteriophage λ DNA* in *RNA polymerase* 413–428 (Cold Spring Harbor Laboratory Press, 1976).
6. Maizels, NM. The nucleotide sequence of the lactose messenger ribonucleic acid transcribed from the UV5 promoter mutant of *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* **70,** 3585–3589 (1973).
7. Kassavetis, G & Chamberlin, M. Pausing and termination of transcription within the early region of bacteriophage T7 DNA in vitro. *J. Biol. Chem.* **256,** 2777–2786 (1981).
8. Reisbig, RR & Hearst, JE. *Escherichia coli* deoxyribonucleic acid-dependent ribonucleic acid polymerase transcriptional pause sites on SV40 DNA F1. *Biochemistry* **20,** 1907–1918 (1981).
9. Herbert, KM *et al.* Sequence-resolved detection of pausing by single RNA polymerase molecules. *Cell* **125,** 1083–1094 (2006).
10. Larson, MH *et al.* A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science* **344,** 1042–1047 (2014).
11. Vvedenskaya, IO *et al.* Interactions between RNA polymerase and the "core recognition element" counteract pausing. *Science* **344,** 1285–1289 (2014).
12. Imashimizu, M *et al.* Visualizing translocation dynamics and nascent transcript errors in paused RNA polymerases in vivo. *Genome Biol.* **16,** 98 (2015).
13. Davenport, RJ, Wuite, GJ, Landick, R & Bustamante, C. Single-molecule study of transcriptional pausing and arrest by *E. coli* RNA polymerase. *Science,* 2497–2500 (2000).
14. Landick, R & Yanofsky, C. Isolation and structural analysis of the *Escherichia coli trp* leader paused transcription complex. *J. Mol. Biol.* **196,** 363–377 (1987).
15. Pan, T, Artsimovitch, I, Fang, XW, Landick, R & Sosnick, TR. Folding of a large ribozyme during transcription and the effect of the elongation factor NusA. *Proc. Natl. Acad. Sci. U.S.A.* **96,** 9545–9550 (1999).
16. Winkler, ME & Yanofsky, C. Pausing of RNA polymerase during in vitro transcription of the tryptophan operon leader region. *Biochemistry* **20,** 3738–3744 (1981).

17. Artsimovitch, I & Landick, R. Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 7090–7095 (2000).

18. Burova, E, Hung, S, Sagitov, V, Stitt, B & Gottesman, M. *Escherichia coli* NusG protein stimulates transcription elongation rates in vivo and in vitro. *J. Bacteriol.* **177,** 1388–1392 (1995).

19. Toulmé, F *et al.* GreA and GreB proteins revive backtracked RNA polymerase in vivo by promoting transcript trimming. *EMBO J.* **19,** 6853–6859 (2000).

20. Weixlbaumer, A, Leon, K, Landick, R & Darst, SA. Structural basis of transcriptional pausing in bacteria. *Cell* **152,** 431–441 (2013).

21. Dangkulwanich, M, Ishibashi, T, Bintu, L & Bustamante, C. Molecular mechanisms of transcription through single-molecule experiments. *Chem. Rev.* **114,** 3203–3223 (2014).

22. Merino, E, Jensen, RA & Yanofsky, C. Evolution of bacterial *trp* operons and their regulation. *Curr. Opin. Microbiol.* **11,** 78–86 (2008).

23. Burns, CM, Richardson, LV & Richardson, JP. Combinatorial effects of NusA and NusG on transcription elongation and rho-dependent termination in *Escherichia coli*. *J. Mol. Biol.* **278,** 307–316 (1998).

24. Lesnik, EA *et al.* Prediction of rho-independent transcriptional terminators in *Escherichia coli*. *Nucleic Acids Res.* **29,** 3583–3594 (2001).

25. Santangelo, TJ & Artsimovitch, I. Termination and antitermination: RNA polymerase runs a stop sign. *Nat. Rev. Microbiol.* **9,** 319 (2011).

26. Ashkin, A. Acceleration and trapping of particles by radiation pressure. *Phys. Rev. Lett.* **24,** 156 (1970).

27. Ashkin, A, Dziedzic, JM, Bjorkholm, J & Chu, S. Observation of a single-beam gradient force optical trap for dielectric particles. *Opt. Lett.* **11,** 288–290 (1986).

28. Ashkin, A & Dziedzic, JM. Optical trapping and manipulation of viruses and bacteria. *Science* **235,** 1517–1520 (1987).

29. Smith, SB, Cui, Y & Bustamante, C. Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science* **271,** 795–799 (1996).

30. Schliwa, M. *Molecular motors* in *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine* 1160–1174 (Springer, 2006).

31. Yin, H *et al.* Transcription against an applied force. *Science* **270,** 1653–1657 (1995).

32. Wuite, GJ, Smith, SB, Young, M, Keller, D & Bustamante, C. Single-molecule studies of the effect of template tension on T7 DNA polymerase activity. *Nature* **404,** 103 (2000).

33. Smith, DE *et al.* The bacteriophage ϕ29 portal motor can package DNA against a large internal force. *Nature* **413,** 748 (2001).

34. Wen, JD *et al.* Following translation by single ribosomes one codon at a time. *Nature* **452,** 598 (2008).

35. Maillard, RA *et al.* ClpX(P) generates mechanical force to unfold and translocate its protein substrates. *Cell* **145,** 459–469 (2011).

36. Mallik, R, Carter, BC, Lex, SA, King, SJ & Gross, SP. Cytoplasmic dynein functions as a gear in response to load. *Nature* **427,** 649 (2004).

37. Block, SM, Goldstein, LS & Schnapp, BJ. Bead movement by single kinesin molecules studied with optical tweezers. *Nature* **348,** 348 (1990).

38. Moffitt, JR, Chemla, YR, Izhaky, D & Bustamante, C. Differential detection of dual traps improves the spatial resolution of optical tweezers. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 9006–9011 (2006).

39. Abbondanzieri, EA, Greenleaf, WJ, Shaevitz, JW, Landick, R & Block, SM. Direct observation of base-pair stepping by RNA polymerase. *Nature* **438,** 460 (2005).

40. Comstock, MJ, Ha, T & Chemla, YR. Ultrahigh-resolution optical trap with single-fluorophore sensitivity. *Nat. Methods* **8,** 335–340 (2011).

41. Dumont, S *et al.* RNA translocation and unwinding mechanism of HCV NS3 helicase and its coordination by ATP. *Nature* **439,** 105 (2006).

42. Moffitt, JR *et al.* Intersubunit coordination in a homomeric ring ATPase. *Nature* **457,** 446 (2009).

43. Sen, M *et al.* The ClpXP protease unfolds substrates using a constant rate of pulling but different gears. *Cell* **155,** 636–646 (2013).

44. Lee, A, Tsekouras, K, Calderon, C, Bustamante, C & Pressé, S. Unraveling the thousand word picture: An introduction to super-resolution data analysis. *Chem. Rev.* **117,** 7276–7330 (2017).

45. Felzenszwalb, PF, Huttenlocher, DP & Kleinberg, JM. *Fast algorithms for large-state-space HMMs with applications to web usage analysis* in *Advances in neural information processing systems* (2004), 409–416.

46. Ma, C, Yang, X & Lewis, PJ. Bacterial transcription as a target for antibacterial drug development. *Microbiol. Mol. Biol. Rev.* **80,** 139–160 (2016).

47. Berg-Sørensen, K & Flyvbjerg, H. Power spectrum analysis for optical tweezers. *Rev. Sci. Instrum.* **75,** 594–612 (2004).

48. Galburt, EA *et al.* Backtracking determines the force sensitivity of RNAP II in a factor-dependent manner. *Nature* **446,** 820 (2007).

49. Toulokhonov, I & Landick, R. The flap domain is required for pause RNA hairpin inhibition of catalysis by RNA polymerase and can modulate intrinsic termination. *Mol. Cell* **12,** 1125–1136 (2003).

50. Dalal, RV *et al.* Pulling on the nascent RNA during transcription does not alter kinetics of elongation or ubiquitous pausing. *Mol. Cell* **23,** 231–239 (2006).

51. Kyzer, S, Ha, KS, Landick, R & Palangat, M. Direct versus limited-step reconstitution reveals key features of an RNA hairpin-stabilized paused transcription complex. *J. Biol. Chem.* **282,** 19020–19028 (2007).

52. Chan, T, Esedoglu, S, Park, F & Yip, A. *Total variation image restoration: Overview and recent developments* in *Handbook of mathematical models in computer vision* 17–31 (Springer, 2006).

53. Johnson, NA. A dynamic programming algorithm for the fused lasso and l_0-segmentation. *J. Comput. Graph. Stat.* **22,** 246–260 (2013).

54. Clauset, A, Shalizi, CR & Newman, ME. Power-law distributions in empirical data. *SIAM Rev.* **51,** 661–703 (2009).

55. Zhou, J, Ha, KS, La Porta, A, Landick, R & Block, SM. Applied force provides insight into transcriptional pausing and its modulation by transcription factor NusA. *Mol. Cell* **44,** 635–646 (2011).

56. Rust, MJ, Bates, M & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3,** 793 (2006).

57. Betzig, E *et al.* Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313,** 1642–1645 (2006).

58. Hess, ST, Girirajan, TP & Mason, MD. Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophys. J.* **91,** 4258–4272 (2006).

59. Sydor, AM, Czymmek, KJ, Puchner, EM & Mennella, V. Super-resolution microscopy: from single molecules to supramolecular assemblies. *Trends Cell Biol.* **25,** 730–748 (2015).

60. Lee, SH, Shin, JY, Lee, A & Bustamante, C. Counting single photoactivatable fluorescent molecules by photoactivated localization microscopy (PALM). *Proc. Natl. Acad. Sci. U.S.A.* **109,** 17436–17441 (2012).

61. Holden, SJ *et al.* Defining the limits of single-molecule FRET resolution in TIRF microscopy. *Biophys. J.* **99,** 3102–3111 (2010).

62. Svoboda, K, Mitra, PP & Block, SM. Fluctuation analysis of motor protein movement and single enzyme kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **91,** 11782–11786 (1994).

63. Chemla, YR, Moffitt, JR & Bustamante, C. Exact solutions for kinetic models of macromolecular dynamics. *J. Phys. Chem. B* **112,** 6025–6044 (2008).

64. Asbury, CL, Fehr, AN & Block, SM. Kinesin moves by an asymmetric hand-over-hand mechanism. *Science* **302,** 2130–2134 (2003).

65. Chistol, G *et al.* High degree of coordination and division of labor among subunits in a homomeric ring ATPase. *Cell* **151,** 1017–1028 (2012).

66. Cordova, JC *et al.* Stochastic but highly coordinated protein unfolding and translocation by the ClpXP proteolytic machine. *Cell* **158,** 647–658 (2014).

67. Cheng, W, Arunajadai, SG, Moffitt, JR, Tinoco, I & Bustamante, C. Single–base pair unwinding and asynchronous RNA release by the hepatitis C virus NS3 helicase. *Science* **333,** 1746–1749 (2011).

68. Callen, HB & Welton, TA. Irreversibility and generalized noise. *Phys. Rev.* **83,** 34 (1951).

69. Finer, JT, Simmons, RM & Spudich, JA. Single myosin molecule mechanics: piconewton forces and nanometre steps. *Nature* **368,** 113–119 (1994).

70. Visscher, K, Gross, SP & Block, SM. Construction of multiple-beam optical traps with nanometer-resolution position sensing. *IEEE J. Sel. Top. Quantum Electron.* **2,** 1066–1076 (1996).

71. Bustamante, C, Chemla, YR & Moffitt, JR. High-resolution dual-trap optical tweezers with differential detection: managing environmental noise. *Cold Spring Harbor Protoc.* **2009,** pdb–ip72 (2009).

72. Bustamante, C, Marko, JF, Siggia, ED, Smith, S, *et al.* Entropic elasticity of lambda-phage DNA. *Science,* 1599–1599 (1994).

73. Neuman, KC & Block, SM. Optical trapping. *Rev. Sci. Instrum.* **75,** 2787–2809 (2004).

74. Mahamdeh, M & Schäffer, E. Optical tweezers with millikelvin precision of temperature-controlled objectives and base-pair resolution. *Opt. Express* **17,** 17190–17199 (2009).

75. Carter, AR, Seol, Y & Perkins, TT. Precision surface-coupled optical-trapping assay with one-basepair resolution. *Biophys. J.* **96,** 2926–2934 (2009).

76. Brau, RR, Tarsa, PB, Ferrer, JM, Lee, P & Lang, MJ. Interlaced optical force-fluorescence measurements for single molecule biophysics. *Biophys. J.* **91,** 1069–1077 (2006).

77. Sasaki, K, Koshioka, M, Misawa, H, Kitamura, N & Masuhara, H. Pattern formation and flow control of fine particles by laser-scanning micromanipulation. *Opt. Lett.* **16,** 1463–1465 (1991).

78. Mio, C, Gong, T, Terray, A & Marr, D. Design of a scanning laser optical trap for multiparticle manipulation. *Rev. Sci. Instrum.* **71,** 2196–2200 (2000).

79. Nambiar, R & Meiners, JC. Fast position measurements with scanning line optical tweezers. *Opt. Lett.* **27,** 836–838 (2002).

80. Valentine, MT *et al.* Precision steering of an optical trap by electro-optic deflection. *Opt. Lett.* **33,** 599–601 (2008).

81. Müllner, FE, Syed, S, Selvin, PR & Sigworth, FJ. Improved hidden Markov models for molecular motors, part 1: basic theory. *Biophys. J.* **99,** 3684–3695 (2010).

82. Syed, S, Müllner, FE, Selvin, PR & Sigworth, FJ. Improved hidden Markov models for molecular motors, part 2: extensions and application to experimental data. *Biophys. J.* **99,** 3696–3703 (2010).

83. Kalafut, B & Visscher, K. An objective, model-independent method for detection of non-uniform steps in noisy signals. *Comput. Phys. Commun.* **179,** 716–723 (2008).

84. Kerssemakers, JW *et al.* Assembly dynamics of microtubules at molecular resolution. *Nature* **442,** 709 (2006).

85. Behnel, S *et al.* Cython: The best of both worlds. *Comput. Sci. Eng.* **13,** 31–39 (2011).

86. Landry, MP, McCall, PM, Qi, Z & Chemla, YR. Characterization of photoactivated singlet oxygen damage in single-molecule optical trap experiments. *Biophys. J.* **97,** 2128–2136 (2009).

87. Komissarova, N & Kashlev, M. RNA polymerase switches between inactivated and activated states by translocating back and forth along the DNA and the RNA. *J. Biol. Chem.* **272,** 15329–15338 (1997).

88. Shaevitz, JW, Abbondanzieri, EA, Landick, R & Block, SM. Backtracking by single RNA polymerase molecules observed at near-base-pair resolution. *Nature* **426,** 684 (2003).

89. Rozovskaya, T, Chenchik, A & Beabealashvilli, RS. Processive pyrophosphorolysis of RNA by *Escherichia coli* RNA polymerase. *FEBS Lett.* **137,** 100–104 (1982).

90. Wang, MD *et al.* Force and velocity measured for single molecules of RNA polymerase. *Science* **282,** 902–907 (1998).

91. Kilchherr, F *et al.* Single-molecule dissection of stacking forces in DNA. *Science* **353,** aaf5508 (2016).

92. Nudler, E. RNA polymerase backtracking in gene regulation and genome instability. *Cell* **149,** 1438–1445 (2012).

93. Epshtein, V, Dutta, D, Wade, J & Nudler, E. An allosteric mechanism of Rho-dependent transcription termination. *Nature* **463,** 245 (2010).

94. Larson, MH, Greenleaf, WJ, Landick, R & Block, SM. Applied force reveals mechanistic and energetic details of transcription termination. *Cell* **132,** 971–982 (2008).

95. Koslover, DJ, Fazal, FM, Mooney, RA, Landick, R & Block, SM. Binding and translocation of termination factor rho studied at the single-molecule level. *J. Mol. Biol.* **423,** 664–676 (2012).

96. Kireeva, ML *et al.* Nucleosome remodeling induced by RNA polymerase II: loss of the H2A/H2B dimer during transcription. *Mol. Cell* **9,** 541–552 (2002).

97. Bintu, L *et al.* The elongation rate of RNA polymerase determines the fate of transcribed nucleosomes. *Nat. Struct. Mol. Biol.* **18,** 1394–1399 (2011).

98. Teves, SS, Weber, CM & Henikoff, S. Transcribing through the nucleosome. *Trends Biochem. Sci.* **39,** 577–586 (2014).

99. Tropea, JE, Cherry, S & Waugh, DS. *Expression and purification of soluble His_6-tagged TEV protease* in *High throughput protein expression and purification* 297–307 (Springer, 2009).

100. Svetlov, V & Artsimovitch, I. *Purification of bacterial RNA polymerase: tools and protocols* in *Bacterial Transcriptional Control* 13–29 (Springer, 2015).

101. Chen, I, Dorr, BM & Liu, DR. A general strategy for the evolution of bond-forming enzymes using yeast display. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 11399–11404 (2011).

102. Lee, J & Borukhov, S. Bacterial RNA polymerase-DNA interaction—the driving force of gene expression and the target for drug action. *Front. Mol. Biosci.* **3,** 73 (2016).

103. Wong, TN, Sosnick, TR & Pan, T. Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 17995–18000 (2007).

104. McGary, K & Nudler, E. RNA polymerase and the ribosome: the close relationship. *Curr. Opin. Microbiol.* **16,** 112–117 (2013).

105. Toulokhonov, I, Zhang, J, Palangat, M & Landick, R. A central role of the RNA polymerase trigger loop in active-site rearrangement during transcriptional pausing. *Mol. Cell* **27,** 406–419 (2007).

106. Zhang, J & Landick, R. A two-way street: Regulatory interplay between RNA polymerase and nascent RNA structure. *Trends Biochem. Sci.* **41,** 293–310 (2016).

107. Komissarova, N & Kashlev, M. Transcriptional arrest: *Escherichia coli* RNA polymerase translocates backward, leaving the 3′ end of the RNA intact and extruded. *Proc. Natl. Acad. Sci. U.S.A.* **94,** 1755–1760 (1997).

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

109. Zamft, B, Bintu, L, Ishibashi, T & Bustamante, C. Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 8948–8953 (2012).

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

111. Kireeva, M *et al.* Millisecond phase kinetic analysis of elongation catalyzed by human, yeast, and *Escherichia coli* RNA polymerase. *Methods* **48,** 333–345 (2009).

112. Larson, MH *et al.* Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 6555–6560 (2012).

113. Comstock, MJ *et al.* Direct observation of structure-function relationship in a nucleic acid–processing enzyme. *Science* **348,** 352–354 (2015).

114. Wales, DJ & Doye, JP. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* **101,** 5111–5116 (1997).

115. Hansen, PC. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Rev.* **34,** 561–580 (1992).

116. Adelman, K *et al.* Single molecule analysis of RNA polymerase elongation reveals uniform kinetic behavior. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 13538–13543 (2002).

117. Sugimoto, N *et al.* Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry* **34,** 11211–11216 (1995).

118. SantaLucia, J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. U.S.A.* **95,** 1460–1465 (1998).

119. Lukačišin, M, Landon, M & Jajoo, R. Sequence-specific thermodynamic properties of nucleic acids influence both transcriptional pausing and backtracking in yeast. *PLoS One* **12,** e0174066 (2017).

120. Depken, M, Galburt, EA & Grill, SW. The origin of short transcriptional pauses. *Biophys. J.* **96,** 2189–2193 (2009).

121. Xayaphoummine, A, Bucher, T & Isambert, H. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res.* **33,** W605–W610 (2005).

122. Chan, CL & Landick, R. Dissection of the *his* leader pause site by base substitution reveals a multipartite signal that includes a pause RNA hairpin. *J. Mol. Biol.* **233,** 25–42 (1993).

123. Wang, D *et al.* Discontinuous movements of DNA and RNA in RNA polymerase accompany formation of a paused transcription complex. *Cell* **81,** 341–350 (1995).

124. Toulokhonov, I, Artsimovitch, I & Landick, R. Allosteric control of RNA polymerase by a site that contacts nascent RNA hairpins. *Science* **292,** 730–733 (2001).

125. Imashimizu, M *et al.* Intrinsic translocation barrier as an initial step in pausing by RNA polymerase II. *J. Mol. Biol.* **425,** 697–712 (2013).

126. Greive, SJ & von Hippel, PH. Thinking quantitatively about transcriptional regulation. *Nat. Rev. Mol. Cell Biol* **6,** 221 (2005).

127. Cheung, AC & Cramer, P. Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature* **471,** 249 (2011).

128. Sekine, Si, Murayama, Y, Svetlov, V, Nudler, E & Yokoyama, S. The ratcheted and ratchetable structural states of RNA polymerase underlie multiple transcriptional functions. *Mol. Cell* **57,** 408–421 (2015).

129. Dangkulwanich, M *et al.* Complete dissection of transcription elongation reveals slow translocation of RNA polymerase II in a linear ratchet mechanism. *eLife* **2** (2013).

130. Maoiléidigh, DÓ, Tadigotla, VR, Nudler, E & Ruckenstein, AE. A unified model of transcription elongation: what have we learned from single-molecule experiments? *Biophys. J.* **100,** 1157–1166 (2011).

131. Kohler, R, Mooney, R, Mills, D, Landick, R & Cramer, P. Architecture of a transcribing-translating expressome. *Science* **356,** 194–197 (2017).

132. Ray-Soni, A, Bellecourt, MJ & Landick, R. Mechanisms of bacterial transcription termination: all good things must end. *Annu. Rev. Biochem.* **85,** 319–347 (2016).

133. Tetone, LE *et al.* Dynamics of GreB-RNA polymerase interaction allow a proofreading accessory protein to patrol for transcription complexes needing rescue. *Proc. Natl. Acad. Sci. U.S.A.* 201616525 (2017).

134. Klar, TA, Jakobs, S, Dyba, M, Egner, A & Hell, SW. Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 8206–8210 (2000).

135. Gustafsson, MG. Nonlinear structured-illumination microscopy: wide-field fluorescence imaging with theoretically unlimited resolution. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 13081–13086 (2005).

136. Wäldchen, S, Lehmann, J, Klein, T, Van De Linde, S & Sauer, M. Light-induced cell damage in live-cell super-resolution microscopy. *Sci. Rep.* **5,** 15348 (2015).

137. Shin, JY *et al.* Visualization and functional dissection of coaxial paired SpoIIIE channels across the sporulation septum. *eLife* **4** (2015).

138. Greenfield, D *et al.* Self-organization of the *Escherichia coli* chemotaxis network imaged with super-resolution light microscopy. *PLoS Biol.* **7,** e1000137 (2009).

139. Pavani, SRP *et al.* Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 2995–2999 (2009).

140. Wells, NP *et al.* Time-resolved three-dimensional molecular tracking in live cells. *Nano Lett.* **10,** 4732–4737 (2010).

141. Thompson, MA, Casolari, JM, Badieirostami, M, Brown, PO & Moerner, W. Three-dimensional tracking of single mRNA particles in *Saccharomyces cerevisiae* using a double-helix point spread function. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 17864–17871 (2010).

142. Ram, S, Kim, D, Ober, RJ & Ward, ES. 3D single molecule tracking with multifocal plane microscopy reveals rapid intercellular transferrin transport at epithelial cell barriers. *Biophys. J.* **103,** 1594–1603 (2012).

143. Welsher, K & Yang, H. Multi-resolution 3D visualization of the early stages of cellular uptake of peptide-coated nanoparticles. *Nat. Nanotechnol.* **9,** 198 (2014).

144. Welsher, K & Yang, H. Imaging the behavior of molecules in biological systems: breaking the 3D speed barrier with 3D multi-resolution microscopy. *Faraday Discuss.* **184,** 359–379 (2015).

145. Li, D *et al.* Extended-resolution structured illumination imaging of endocytic and cytoskeletal dynamics. *Science* **349,** aab3500 (2015).

146. Shechtman, Y, Weiss, LE, Backer, AS, Sahl, SJ & Moerner, W. Precise three-dimensional scan-free multiple-particle tracking over large axial ranges with tetrapod point spread functions. *Nano Lett.* **15,** 4194–4199 (2015).

147. Lippincott-Schwartz, J. Profile of Eric Betzig, Stefan Hell, and WE Moerner, 2014 Nobel Laureates in Chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **112,** 2630–2632 (2015).

148. Balzarotti, F *et al.* Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. *Science,* aak9913 (2016).

149. Gunawardena, J. Models in biology:'accurate descriptions of our pathetic thinking'. *BMC Biol.* **12,** 29 (2014).

150. Rollins, GC, Shin, JY, Bustamante, C & Pressé, S. Stochastic approach to the molecular counting problem in superresolution microscopy. *Proc. Natl. Acad. Sci. U.S.A.* **112,** E110–E118 (2015).

151. Berglund, AJ. Statistics of camera-based single-particle tracking. *Phys. Rev. E* **82,** 011917 (2010).

152. De Chaumont, F *et al.* Icy: an open bioimage informatics platform for extended reproducible research. *Nat. Methods* **9,** 690 (2012).

153. Lindén, M, Ćurić, V, Boucharin, A, Fange, D & Elf, J. Simulated single molecule microscopy with SMeagol. *Bioinformatics* **32,** 2394–2395 (2016).

154. Calderon, CP. Motion blur filtering: a statistical approach for extracting confinement forces and diffusivity from a single blurred trajectory. *Phys. Rev. E* **93,** 053303 (2016).

155. Masters, BR. *The Development of Fluorescence Microscopy* in *Encyclopedia of Life Sciences* 1–9 (John Wiley & Sons, 2010).

156. Coons, AH, Creech, HJ & Jones, RN. Immunological properties of an antibody containing a fluorescent group. *Proc. Soc. Exp. Biol. Med.* **47,** 200–202 (1941).

157. Shimomura, O, Johnson, FH & Saiga, Y. Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, *Aequorea*. *J. Cell. Physiol.* **59,** 223–239 (1962).

158. Tsien, RY. The green fluorescent protein. *Annu. Rev. Biochem.* **67,** 509–544 (1998).

159. Moerner, WE & Kador, L. Optical detection and spectroscopy of single molecules in a solid. *Phys. Rev. Lett.* **62,** 2535 (1989).

160. Shera, EB, Seitzinger, NK, Davis, LM, Keller, RA & Soper, SA. Detection of single fluorescent molecules. *Chem. Phys. Lett.* **174,** 553–557 (1990).

161. Airy, G. On the Diffraction of an Object-glass with a Circular Aperture. *Trans. Camb. Phil. Soc.* **5,** 283–291 (1834).

162. Abbe, E. Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung. *Archiv für mikroskopische Anatomie* **9,** 413–418 (1873).

163. Rayleigh. XXXI. Investigations in optics, with special reference to the spectroscope. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **8,** 261–274 (1879).

164. Hecht, E. *Optics* 4th edition (Addison-Wesley, 2002).

165. Izeddin, I *et al.* PSF shaping using adaptive optics for three-dimensional single-molecule super-resolution imaging and tracking. *Opt. Express* **20,** 4957–4967 (2012).

166. Holden, SJ, Uphoff, S & Kapanidis, AN. DAOSTORM: an algorithm for high-density super-resolution microscopy. *Nat. Methods* **8,** 279 (2011).

167. Kourkoutis, LF, Plitzko, JM & Baumeister, W. Electron microscopy of biological materials at the nanometer scale. *Annu. Rev. Mater. Res.* **42,** 33–58 (2012).

168. Carroll, RJ & Hall, P. Optimal rates of convergence for deconvolving a density. *J. Am. Stat. Assoc.* **83,** 1184–1186 (1988).

169. Fan, J. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Stat.* 1257–1272 (1991).

170. Agard, DA & Sedat, JW. Three-dimensional architecture of a polytene nucleus. *Nature* **302,** 676 (1983).

171. Schermelleh, L, Heintzmann, R & Leonhardt, H. A guide to super-resolution fluorescence microscopy. *J. Cell Biol.* **190,** 165–175 (2010).

172. Morrisson, IE, Anderson, CM, Georgiou, GN & Cherry, RJ. Measuring diffusion coefficients of labelled particles on cell surfaces by digital fluorescence microscopy. *Biochem. Soc. Trans.* **18,** 938 (1990).

173. Anderson, CM, Georgiou, GN, Morrison, IE, Stevenson, G & Cherry, RJ. Tracking of cell surface receptors by fluorescence digital imaging microscopy using a charge-coupled device camera. Low-density lipoprotein and influenza virus receptor mobility at 4 °C. *J. Cell Sci.* **101,** 415–425 (1992).

174. Betzig, E. Proposed method for molecular optical imaging. *Opt. Lett.* **20,** 237–239 (1995).

175. Ando, R, Hama, H, Yamamoto-Hino, M, Mizuno, H & Miyawaki, A. An optical marker based on the UV-induced green-to-red photoconversion of a fluorescent protein. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 12651–12656 (2002).

176. Bates, M, Blosser, TR & Zhuang, X. Short-range spectroscopic ruler based on a single-molecule optical switch. *Phys. Rev. Lett.* **94,** 108101 (2005).

177. Hell, SW & Wichmann, J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Opt. Lett.* **19,** 780–782 (1994).

178. Jost, A & Heintzmann, R. Superresolution multidimensional imaging with structured illumination microscopy. *Ann. Rev. Mater. Res.* **43,** 261–282 (2013).

179. Shroff, H *et al.* Dual-color superresolution imaging of genetically expressed probes within individual adhesion complexes. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 20308–20313 (2007).

180. Wang, S, Moffitt, JR, Dempsey, GT, Xie, XS & Zhuang, X. Characterization and development of photoactivatable fluorescent proteins for single-molecule–based superresolution imaging. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 8452–8457 (2014).

181. Lidke, KA, Rieger, B, Jovin, TM & Heintzmann, R. Superresolution by localization of quantum dots using blinking statistics. *Opt. Express* **13,** 7052–7062 (2005).

182. Alivisatos, AP, Gu, W & Larabell, C. Quantum dots as cellular probes. *Annu. Rev. Biomed. Eng.* **7,** 55–76 (2005).

183. Press, W, Teukolsky, S, Vetterling, W & Flannery, B. *Numerical Recipes 3rd Edition: The Art of Scientific Computing* (Cambridge University Press, 2007).

184. Huang, ZL *et al.* Localization-based super-resolution microscopy with an sCMOS camera. *Opt. Express* **19,** 19156–19168 (2011).

185. Hynecek, J & Nishiwaki, T. Excess noise and other important characteristics of low light level imaging using charge multiplying CCDs. *IEEE Trans. Electron Devices* **50,** 239–245 (2003).

186. Ulbrich, MH & Isacoff, EY. Subunit counting in membrane-bound proteins. *Nat. Methods* **4,** 319 (2007).

187. Pertsinidis, A, Zhang, Y & Chu, S. Subnanometre single-molecule localization, registration and distance measurements. *Nature* **466,** 647 (2010).

188. Huang, F *et al.* Video-rate nanoscopy using sCMOS camera–specific single-molecule localization algorithms. *Nat. Methods* **10,** 653 (2013).

189. Backlund, MP, Lew, MD, Backer, AS, Sahl, SJ & Moerner, W. The role of molecular dipole orientation in single-molecule fluorescence microscopy and implications for super-resolution imaging. *ChemPhysChem* **15,** 587–599 (2014).

190. Rowland, DJ & Biteen, JS. Top-hat and asymmetric gaussian-based fitting functions for quantifying directional single-molecule motion. *ChemPhysChem* **15,** 712–720 (2014).

191. Ghosh, RN & Webb, WW. Automated detection and tracking of individual and clustered cell surface low density lipoprotein receptor molecules. *Biophys. J.* **66,** 1301–1318 (1994).

192. Sage, D *et al.* Quantitative evaluation of software packages for single-molecule localization microscopy. *Nat. Methods* **12,** 717 (2015).

193. Sage, D, Neumann, FR, Hediger, F, Gasser, SM & Unser, M. Automatic tracking of individual fluorescence particles: application to the study of chromosome dynamics. *IEEE Trans. Image Process.* **14,** 1372–1383 (2005).

194. Köthe, U, Herrmannsdörfer, F, Kats, I & Hamprecht, FA. SimpleSTORM: a fast, self-calibrating reconstruction algorithm for localization microscopy. *Histochem. Cell Biol.* **141,** 613–627 (2014).

195. Turin, G. An introduction to matched filters. *IRE Trans. Inf. Theory* **6,** 311–329 (1960).

196. Coltharp, C, Kessler, RP & Xiao, J. Accurate construction of photoactivated localization microscopy (PALM) images for quantitative measurements. *PLoS One* **7,** e51725 (2012).

197. Křížek, P, Raška, I & Hagen, GM. Minimizing detection errors in single molecule localization microscopy. *Opt. Express* **19,** 3226–3235 (2011).

198. Rao, CR. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37,** 81–91 (1945).

199. Cramér, H. *Mathematical Methods of Statistics* (Princeton University Press, 1946).

200. Ober, RJ, Ram, S & Ward, ES. Localization accuracy in single-molecule microscopy. *Biophys. J.* **86,** 1185–1200 (2004).

201. Abraham, AV, Ram, S, Chao, J, Ward, E & Ober, RJ. Quantitative study of single molecule location estimation techniques. *Opt. Express* **17,** 23352–23373 (2009).

202. Mortensen, KI, Churchman, LS, Spudich, JA & Flyvbjerg, H. Optimized localization analysis for single-molecule tracking and super-resolution microscopy. *Nat. Methods* **7,** 377 (2010).

203. Smith, CS, Joseph, N, Rieger, B & Lidke, KA. Fast, single-molecule localization that achieves theoretically minimum uncertainty. *Nat. Methods* **7,** 373 (2010).

204. Starr, R, Stahlheber, S & Small, A. Fast maximum likelihood algorithm for localization of fluorescent molecules. *Opt. Lett.* **37,** 413–415 (2012).

205. Nieuwenhuizen, RP *et al.* Measuring image resolution in optical nanoscopy. *Nat. Methods* **10,** 557 (2013).

206. Banterle, N, Bui, KH, Lemke, EA & Beck, M. Fourier ring correlation as a resolution criterion for super-resolution microscopy. *J. Struct. Biol.* **183,** 363–367 (2013).

207. Liu, SL *et al.* Fast and high-accuracy localization for three-dimensional single-particle tracking. *Sci. Rep.* **3,** 2462 (2013).

208. Cheezum, MK, Walker, WF & Guilford, WH. Quantitative comparison of algorithms for tracking single fluorescent particles. *Biophys. J.* **81,** 2378–2388 (2001).

209. Parthasarathy, R. Rapid, accurate particle tracking by calculation of radial symmetry centers. *Nat. Methods* **9,** 724 (2012).

210. Ma, H, Long, F, Zeng, S & Huang, ZL. Fast and precise algorithm based on maximum radial symmetry for single molecule localization. *Opt. Lett.* **37,** 2481–2483 (2012).

211. Guizar-Sicairos, M, Thurman, ST & Fienup, JR. Efficient subpixel image registration algorithms. *Opt. Lett.* **33,** 156–158 (2008).

212. York, AG, Ghitani, A, Vaziri, A, Davidson, MW & Shroff, H. Confined activation and subdiffractive localization enables whole-cell PALM with genetically expressed probes. *Nat. Methods* **8,** 327 (2011).

213. Hess, ST *et al.* Dynamic clustered distribution of hemagglutinin resolved at 40 nm in living cell membranes discriminates between raft theories. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 17370–17375 (2007).

214. Thompson, RE, Larson, DR & Webb, WW. Precise nanometer localization analysis for individual fluorescent probes. *Biophys. J.* **82,** 2775–2783 (2002).

215. Yildiz, A *et al.* Myosin V walks hand-over-hand: single fluorophore imaging with 1.5-nm localization. *Science* **300,** 2061–2065 (2003).

216. Anscombe, FJ. The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35,** 246–254 (1948).

217. Murtagh, F, Starck, JL & Bijaoui, A. Image restoration with noise suppression using a multiresolution support. *Astron. Astrophys. Suppl. Ser.* **112,** 179 (1995).

218. Stallinga, S & Rieger, B. Accuracy of the Gaussian point spread function model in 2D localization microscopy. *Opt. Express* **18,** 24461–24476 (2010).

219. Vaughan, JC, Jia, S & Zhuang, X. Ultrabright photoactivatable fluorophores created by reductive caging. *Nat. Methods* **9,** 1181 (2012).

220. Engelhardt, J *et al.* Molecular orientation affects localization accuracy in superresolution far-field fluorescence microscopy. *Nano Lett.* **11,** 209–213 (2010).

221. Zipfel, WR, Williams, RM & Webb, WW. Nonlinear magic: multiphoton microscopy in the biosciences. *Nat. Biotechnol.* **21,** 1369 (2003).

222. Huisken, J, Swoger, J, Del Bene, F, Wittbrodt, J & Stelzer, EH. Optical sectioning deep inside live embryos by selective plane illumination microscopy. *Science* **305,** 1007–1009 (2004).

223. Kao, HP & Verkman, A. Tracking of single fluorescent particles in three dimensions: use of cylindrical optics to encode particle position. *Biophys. J.* **67,** 1291–1300 (1994).

224. Huang, B, Wang, W, Bates, M & Zhuang, X. Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. *Science* **319,** 810–813 (2008).

225. Juette, MF *et al.* Three-dimensional sub–100 nm resolution fluorescence microscopy of thick samples. *Nat. Methods* **5,** 527 (2008).

226. Shtengel, G *et al.* Interferometric fluorescent super-resolution microscopy resolves 3D cellular ultrastructure. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 3125–3130 (2009).

227. Stetson, PB. DAOPHOT: A computer program for crowded-field stellar photometry. *Publ. Astron. Soc. Pac.* **99,** 191 (1987).

228. Huang, F, Schwartz, SL, Byars, JM & Lidke, KA. Simultaneous multiple-emitter fitting for single molecule super-resolution imaging. *Biomed. Opt. Express* **2,** 1377–1393 (2011).

229. Wilks, SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9,** 60–62 (1938).

230. Candes, EJ, Romberg, JK & Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59,** 1207–1223 (2006).

231. Zhu, L, Zhang, W, Elnatan, D & Huang, B. Faster STORM using compressed sensing. *Nat. Methods* **9,** 721 (2012).

232. Grant, M, Boyd, S & Ye, Y. *Disciplined convex programming* in *Global optimization* 155–210 (Springer, 2006).

233. Burnette, DT, Sengupta, P, Dai, Y, Lippincott-Schwartz, J & Kachar, B. Bleaching/blinking assisted localization microscopy for superresolution imaging using standard fluorescent molecules. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 21081–21086 (2011).

234. Dertinger, T, Colyer, R, Iyer, G, Weiss, S & Enderlein, J. Fast, background-free, 3D super-resolution optical fluctuation imaging (SOFI). *Proc. Natl. Acad. Sci. U.S.A.* **106,** 22287–22292 (2009).

235. Hyvarinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* **10,** 626–634 (1999).

236. Cox, S *et al.* Bayesian localization microscopy reveals nanoscale podosome dynamics. *Nat. Methods* **9,** 195 (2012).

237. Hu, YS, Nan, X, Sengupta, P, Lippincott-Schwartz, J & Cang, H. Accelerating 3B single-molecule super-resolution microscopy with cloud computing. *Nat. Methods* **10,** 96 (2013).

238. Mukamel, EA, Babcock, H & Zhuang, X. Statistical deconvolution for superresolution fluorescence microscopy. *Biophys. J.* **102,** 2391–2400 (2012).

239. Min, J *et al.* FALCON: fast and unbiased reconstruction of high-density super-resolution microscopy data. *Sci. Rep.* **4,** 4577 (2014).

240. Bates, M, Huang, B, Dempsey, GT & Zhuang, X. Multicolor super-resolution imaging with photo-switchable fluorescent probes. *Science* **317,** 1749–1753 (2007).

241. Mlodzianoski, MJ *et al.* Sample drift correction in 3D fluorescence photoactivation localization microscopy. *Opt. Express* **19,** 15009–15019 (2011).

242. McGorty, R, Kamiyama, D & Huang, B. Active microscope stabilization in three dimensions using image correlation. *Opt. Nanoscopy* **2,** 3 (2013).

243. Nieuwenhuizen, RP *et al.* Quantitative localization microscopy: effects of photophysics and labeling stoichiometry. *PLoS One* **10,** e0127989 (2015).

244. Berg, HC. The rotary motor of bacterial flagella. *Annu. Rev. Biochem.* **72** (2003).

245. Leake, MC *et al.* Stoichiometry and turnover in single, functioning membrane protein complexes. *Nature* **443,** 355 (2006).

246. Delalez, NJ *et al.* Signal-dependent turnover of the bacterial flagellar switch protein FliM. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 11347–11351 (2010).

247. Wu, LJ, Lewis, PJ, Allmansberger, R, Hauser, PM & Errington, J. A conjugation-like mechanism for prespore chromosome partitioning during sporulation in *Bacillus subtilis*. *Genes Dev.* **9,** 1316–1326 (1995).

248. Liu, NJL, Dutton, RJ & Pogliano, K. Evidence that the SpoIIIE DNA translocase participates in membrane fusion during cytokinesis and engulfment. *Mol. Microbiol.* **59,** 1097–1113 (2006).

249. Baker, MD, Wolanin, PM & Stock, JB. Signal transduction in bacterial chemotaxis. *Bioessays* **28,** 9–22 (2006).

250. Gross, D & Webb, WW. Molecular counting of low-density lipoprotein particles as individuals and small clusters on cell surfaces. *Biophys. J.* **49,** 901–911 (1986).

251. Burton, BM, Marquis, KA, Sullivan, NL, Rapoport, TA & Rudner, DZ. The ATPase SpoIIIE transports DNA across fused septal membranes during sporulation in *Bacillus subtilis*. *Cell* **131,** 1301–1312 (2007).

252. Joglekar, AP, Bouck, DC, Molk, JN, Bloom, KS & Salmon, ED. Molecular architecture of a kinetochore–microtubule attachment site. *Nat. Cell Biol.* **8,** 581 (2006).

253. Coffman, VC, Wu, P, Parthun, MR & Wu, JQ. CENP-A exceeds microtubule attachment sites in centromere clusters of both budding and fission yeast. *J. Cell Biol.* **195,** 563–572 (2011).

254. Lawrimore, J, Bloom, KS & Salmon, E. Point centromeres contain more than a single centromere-specific Cse4 (CENP-A) nucleosome. *J. Cell Biol.* **195,** 573–582 (2011).

255. Henikoff, S & Henikoff, JG. "Point" centromeres of Saccharomyces harbor single centromere-specific nucleosomes. *Genetics* **190,** 1575–1577 (2012).

256. McKinley, KL & Cheeseman, IM. The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol.* **17,** 16 (2016).

257. Liesche, C *et al.* Automated analysis of single-molecule photobleaching data by statistical modeling of spot populations. *Biophys. J.* **109,** 2352–2362 (2015).

258. Ambrose, WP *et al.* Fluorescence photon antibunching from single molecules on a surface. *Chem. Phys. Lett.* **269,** 365–370 (1997).

259. Kurz, A *et al.* Counting fluorescent dye molecules on DNA origami by means of photon statistics. *Small* **9,** 4061–4068 (2013).

260. Messina, TC, Kim, H, Giurleo, JT & Talaga, DS. Hidden Markov model analysis of multichromophore photobleaching. *J. Phys. Chem. B* **110,** 16366–16376 (2006).

261. Rabiner, LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77,** 257–286 (1989).

262. Andrec, M, Levy, RM & Talaga, DS. Direct determination of kinetic rates from single-molecule photon arrival trajectories using hidden Markov models. *J. Phys. Chem. A* **107,** 7454–7464 (2003).

263. McKinney, SA, Joo, C & Ha, T. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys. J.* **91,** 1941–1951 (2006).

264. McGuire, H, Aurousseau, MR, Bowie, D & Blunck, R. Automating single subunit counting of membrane proteins in mammalian cells. *J. Biol. Chem.* **287,** 35912–35921 (2012).

265. Chen, Y, Deffenbaugh, NC, Anderson, CT & Hancock, WO. Molecular counting by photobleaching in protein complexes with many subunits: best practices and application to the cellulose synthesis complex. *Mol. Biol. Cell* **25,** 3630–3642 (2014).

266. Tsekouras, K, Custer, TC, Jashnsaz, H, Walter, NG & Pressé, S. A novel method to accurately locate and count large numbers of steps by photobleaching. *Mol. Biol. Cell* **27,** 3601–3615 (2016).

267. Carter, BC, Vershinin, M & Gross, SP. A comparison of step-detection methods: how well can you do? *Biophys. J.* **94,** 306–319 (2008).

268. Chung, S & Kennedy, R. Forward-backward non-linear filtering technique for extracting small biological signals from noise. *J. Neurosci. Methods* **40,** 71–86 (1991).

269. Killick, R, Fearnhead, P & Eckley, IA. Optimal detection of changepoints with a linear computational cost. *J. Am. Stat. Assoc.* **107,** 1590–1598 (2012).

270. Rigaill, G. A pruned dynamic programming algorithm to recover the best segmentations with 1 to $K_{\max}$ change-points. *J. Soc. Stat. Paris* **156,** 180–205 (2015).

271. Schwarz, G *et al.* Estimating the dimension of a model. *Ann. Stat.* **6,** 461–464 (1978).

272. Annibale, P, Scarselli, M, Kodiyan, A & Radenovic, A. Photoactivatable fluorescent protein mEos2 displays repeated photoactivation after a long-lived dark state in the red photoconverted form. *J. Phys. Chem. Lett.* **1,** 1506–1510 (2010).

273. Dickson, RM, Cubitt, AB, Tsien, RY & Moerner, W. On/off blinking and switching behaviour of single molecules of green fluorescent protein. *Nature* **388,** 355 (1997).

274. Shaner, NC, Steinbach, PA & Tsien, RY. A guide to choosing fluorescent proteins. *Nat. Methods* **2,** 905 (2005).

275. Durisic, N, Laparra-Cuervo, L, Sandoval-Álvarez, Á, Borbely, JS & Lakadamyali, M. Single-molecule evaluation of fluorescent protein photoactivation efficiency using an in vivo nanotemplate. *Nat. Methods* **11,** 156 (2014).

276. Bálint, Š, Vilanova, IV, Álvarez, ÁS & Lakadamyali, M. Correlative live-cell and super-resolution microscopy reveals cargo transport dynamics at microtubule intersections. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 3375–3380 (2013).

277. Rosenbloom, AB *et al.* Optimized two-color super resolution imaging of Drp1 during mitochondrial fission with a slow-switching Dronpa variant. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 13093–13098 (2014).

278. Löschberger, A *et al.* Super-resolution imaging visualizes the eightfold symmetry of gp210 proteins around the nuclear pore complex and resolves the central channel with nanometer resolution. *J Cell Sci* **125,** 570–575 (2012).

279. Szymborska, A *et al.* Nuclear pore scaffold structure analyzed by super-resolution microscopy and particle averaging. *Science* **341,** 655–658 (2013).

280. Tavakoli, M, Taylor, JN, Li, CB, Komatsuzaki, T & Pressé, S. Single molecule data analysis: an introduction. *arXiv preprint arXiv:1606.00403* (2016).

281. Ferguson, TS. A Bayesian analysis of some nonparametric problems. *Ann. Stat.* 209–230 (1973).

282. Chen, BC *et al.* Lattice light-sheet microscopy: imaging molecules to embryos at high spatiotemporal resolution. *Science* **346,** 1257998 (2014).

283. Bartsch, TF, Kochanczyk, MD, Lissek, EN, Lange, JR & Florin, EL. Nanoscopic imaging of thick heterogeneous soft-matter structures in aqueous solution. *Nat. Commun.* **7,** 12729 (2016).

284. Van de Linde, S, Wolter, S, Heilemann, M & Sauer, M. The effect of photoswitching kinetics and labeling densities on super-resolution fluorescence imaging. *J. Biotechnol.* **149,** 260–266 (2010).

285. Ha, T & Tinnefeld, P. Photophysics of fluorescent probes for single-molecule biophysics and super-resolution imaging. *Annu. Rev. Phys. Chem.* **63,** 595–617 (2012).

286. Masson, JB *et al.* Mapping the energy and diffusion landscapes of membrane proteins at the cell surface using high-density single-molecule imaging and Bayesian inference: application to the multiscale dynamics of glycine receptors in the neuronal membrane. *Biophys. J.* **106,** 74–83 (2014).

287. Chertkov, M, Kroc, L, Krzakala, F, Vergassola, M & Zdeborová, L. Inference in particle tracking experiments by passing messages between images. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 7663–7668 (2010).

288. Chenouard, N, Bloch, I & Olivo-Marin, JC. Multiple hypothesis tracking for cluttered biological image sequences. *IEEE Trans. Pattern Anal. Mach. Intell.* **35,** 2736–3750 (2013).

289. Teh, Y, Jordan, M, Beal, M & Blei, D. Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.* **101,** 1566–1581 (2006).

290. Calderon, CP & Bloom, K. Inferring latent states and refining force estimates via hierarchical Dirichlet process modeling in single particle tracking experiments. *PLoS One* **10,** e0137633 (2015).

291. Ghosh, I, Hamilton, AD & Regan, L. Antiparallel leucine zipper-directed protein reassembly: application to the green fluorescent protein. *J. Am. Chem. Soc.* **122,** 5658–5659 (2000).

292. Hu, CD, Chinenov, Y & Kerppola, TK. Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation. *Mol. Cell* **9,** 789–798 (2002).

293. Lindman, S, Hernandez-Garcia, A, Szczepankiewicz, O, Frohm, B & Linse, S. In vivo protein stabilization based on fragment complementation and a split GFP system. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 19826–19831 (2010).

294. Lalkens, B, Testa, I, Willig, KI & Hell, SW. Nanoscopy of protein colocalization in living cells by STED and GSDIM. *Microsc. Res. Tech.* **75,** 1–6 (2012).

295. Lee, YR *et al.* Development of bimolecular fluorescence complementation using Dronpa for visualization of protein–protein interactions in cells. *Mol. Imaging Biol.* **12,** 468–478 (2010).

296. Habuchi, S *et al.* Reversible single-molecule photoswitching in the GFP-like fluorescent protein Dronpa. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 9511–9516 (2005).

297. Gurskaya, NG *et al.* Engineering of a monomeric green-to-red photoactivatable fluorescent protein induced by blue light. *Nat. Biotechnol.* **24,** 461 (2006).

298. Subach, FV *et al.* Photoactivatable mCherry for high-resolution two-color fluorescence microscopy. *Nat. Methods* **6,** 153 (2009).

299. Gavin, PD, Devenish, RJ & Prescott, M. FRET reveals changes in the F_1–stator stalk interaction during activity of F_1F_0-ATP synthase. *Biochim. Biophys. Acta – Bioenergetics* **1607,** 167–179 (2003).

300. von Ballmoos, C, Wiedenmann, A & Dimroth, P. Essentials for ATP synthesis by F_1F_0 ATP synthases. *Annu. Rev. Biochem.* **78,** 649–672 (2009).

301. Rizzuto, R, Simpson, AW, Brini, M & Pozzan, T. Rapid changes of mitochondrial Ca^{2+} revealed by specifically targeted recombinant aequorin. *Nature* **358,** 325 (1992).

302. Fan, JY *et al.* Split mCherry as a new red bimolecular fluorescence complementation system for visualizing protein–protein interactions in living cells. *Biochem. Biophys. Res. Commun.* **367,** 47–53 (2008).

303. Nickerson, A, Huang, T, Lin, LJ & Nan, X. Photoactivated localization microscopy with bimolecular fluorescence complementation (BiFC-PALM) for nanoscale imaging of protein-protein interactions in cells. *PLoS One* **9,** e100589 (2014).

304. Liu, Z *et al.* Super-resolution imaging and tracking of protein–protein interactions in sub-diffraction cellular space. *Nat. Commun.* **5,** 4443 (2014).