Untangling reaction pathways through modern approaches to high-throughput single-molecule force-spectroscopy experiments

David Dulin², Bojk A Berghuis¹, Martin Depken¹ and Nynke H Dekker¹

Single-molecule experiments provide a unique means for real-time observation of the activity of individual biomolecular machines. Through such techniques, insights into the mechanics of for example, polymerases, helicases, and packaging motors have been gleaned. Here we describe the recent advances in single-molecule force spectroscopy instrumentation that have facilitated high-throughput acquisition at high spatiotemporal resolution. The large datasets attained by such methods can capture rare but important events, and contain information regarding stochastic behaviors covering many orders of magnitude in time. We further discuss analysis of such data sets, and with a special focus on the pause states described in the general literature on RNA polymerase pausing we compare and contrast the signatures of different reaction pathways.

Addresses
¹ Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology, Lorentzweg 1, 2628 CJ Delft, The Netherlands
² Department of Physics, Clarendon Laboratory, University of Oxford, Parks Road, Oxford OX1 3PU, United Kingdom

Corresponding author: Dekker, Nynke H (n.h.dekker@tudelft.nl)

Introduction
The correct readout, maintenance, repair, and replication of genomic information involves a stunning variety of carefully coordinated and regulated molecular machines, including key players such as RNA polymerases (RNAP) and replisomes. The progression of these machines is highly dynamic, as their activity is frequently and stochastically interrupted by numerous co-factors or intrinsic catalytic events. For example, during transcription, RNAP progression is interrupted by various types of pauses, including backtracking pauses resulting from RNAP diffusing upstream on the DNA template [1,2], and regulatory pauses allowing for transcription-translation synchronization and RNA co-transcriptional folding [3,4*,5]. During replication, the replisome also frequently halts, for example, for the formation of primers on the Okazaki fragment [6], due to DNA polymerase exchange [7], and possibly also due to proofreading [8–10].

To obtain quantitative insight into the functioning of molecular machines, probing at the single-molecule level with nanometre-scale spatial and millisecond temporal resolution has proved to be very successful [11]. This success results from the ability to detect transient intermediates or rare events that are masked when ensemble techniques are used. Single-molecule methods may be broadly grouped into optical methods visualizing individual molecules using fluorescence-based microscopy [12–14], and force-based methods such as atomic force microscopy, flow-based stretching, and optical and magnetic tweezers [15]. While both approaches have provided invaluable insight into the functioning of molecular machines, we here focus on the latter. Motivated by the sub-nanometer elementary length scales of biological substrates such as DNA, much effort has gone into developing instruments with high spatiotemporal resolution [11]. Provided sufficiently low enzyme kinetic rates (~1 bp/s), optical tweezers have been developed with single DNA base pair (bp) resolution [16]. Such tweezers have allowed detailed mechanistic studies of, for example, RNAP [17*], the Φ29 packaging motor [18*], and the Hepatitis C viral helicase NS3 [19].

Though biological systems that catalyze chemical reactions along a single pathway [18*,20,21,42*] have successfully been characterized with high-resolution optical tweezers, approaches suitable for analysis of more complicated pathways displayed remain a challenge due to the low yield of such technique. We here review recent approaches to parallelized single-molecule force-spectroscopy methods using widefield (e.g. camera-based) detection to simultaneously track the action of a large number of independent molecules and substrates (Figure 1). For the case of polymerizing enzymes with an internal pause dynamics, we further discuss an analysis approach that capitalizes on large data sets to reveal information about the enzyme translocation pathway, and review the signatures of various pause types described in the literature.
Single-molecule approaches go parallel

The simplest high-throughput single-molecule approach, typically denoted flow-stretch experiments, involves the use of liquid flow to exert drag forces on tethered beads (Figure 1a). Flow-stretch experiments have the advantage of being relatively simple while still allowing for the observation of enzymatic activity on long DNA templates (>40 kb) at about a 100 nm resolution. This approach has been extensively used to study the dynamics of replication in the bacteriophage T7 model system [22,23]. A more recent approach, denoted nanophotonic standing-wave array trap (nSWAT), uses microfluidics combined with waveguiding to create an array of optical traps [24] (Figure 1b). Being integrated into a micro-fabricated chip, the nSWAT assay substantially reduces the complexity of optical tweezers and the influence of mechanical noise. Another recent approach is acoustic force spectroscopy (AFS), which uses a flow-cell integrated piezo element to apply acoustic pressure to micron-sized beads [25] (Figure 1c). AFS is ideal for high-throughput experiments, as it generates a homogenous force over distances much larger than the sizes of the force transducers. For the same reasons, magnetic tweezers (MT) are also inherently suitable for parallelization [26*] (Figure 1d). The capabilities of a widefield camera-based imaging approach combined with MT were first demonstrated by following either tens of molecules in real-time [27], or hundreds with post-processing image analysis [28]. The potential for multiplexing is further expanded by a recent increase in camera resolution (Figure 1d), and the use of graphics processing units interfaced with the CUDA-based parallel computing framework [26*,41*,46]. It is now possible to simultaneously follow up to 800 beads in real-time, with a resolution of 1 nm at 25 Hz [26*].
Further development of high-throughput experiments will greatly help the mapping of single-molecule events that are either very long-lived or rare. We now discuss how to visualize such data, and illustrate the two main scenario where multiplexing becomes important by discussing two recent studies utilizing high-throughput MT approaches to study DNA binding proteins and polymerase activity.

**Dwell-time distributions**

Dwell-time analysis has a long history of being applied to single-molecule experiments where it is possible to directly measure the duration (the dwell-time) of the event of interest [29,42]. With the increase in the type of events that can be studied with statistical significance in multiplexed force-spectroscopy experiments, it becomes important to be able to visualize the data over many orders of magnitude in time and probability. For such visualizations it is especially useful to collect dwell times in a weighted histogram, where the score in each bin is normalized with the bin width and total number of dwell times recorded. Such weighted histogram represents empirical dwell-time distributions (DTD). The basic shape of an empirical DTD is insensitive to the binning used, which makes it possible to consistently visualize the data over several orders of magnitude by using various binning schemes (e.g. using log-scaled bins).

**Magnetic tweezers-based study of DNA binding proteins capitalize on improved statistics**

High-throughput proved essential in a recent study probing *Escherichia coli* replication fork arrest induced by the tight interaction between the DNA binding protein Tus and its cognate DNA-binding sequence *Ter* [30**]. It had previously been argued that the tight interaction was dependent on specific interactions between Tus and the replisome [31]. By using a multiplexed MT DNA hairpin assay (Figure 2a) to mimicking fork progression without the proteins of the replisome, it was shown that protein–protein interactions were not necessary for tight Tus–Ter interactions. Gathering the necessary dwell-time statistics would have been very demanding in time without multiplexing, as strand separation frequently remained blocked at the Tus–Ter site for hundreds of seconds (Figure 2b).

**Magnetic tweezers-based study of polymerase activity capitalize on improved statistics**

A recent study of viral mutagenesis used high-throughput magnetic tweezers to examine rare nucleotide misincorporation events during replication [32**] (Figure 1d). In this work, the progression of a viral RNA-dependent RNA polymerase (RdRp) was monitored at the single-molecule level.

---

**Figure 2**

High-throughput single-molecule assays and the resulting multiplexed datasets. (a) In the Tus–Ter experimental assay, the DNA–protein lock is formed by force-induced DNA hairpin strand separation. The interaction strength is quantified by measuring the lifetime of the lock at a constant force. (b) This yields a dataset of dwell-times until rupture events (i.e., full opening of the hairpin). The resulting dwell-time distribution shows evidence of multiple exponential states, revealing the intermediate steps towards lock formation (inset). (c) A pair of magnets is used to apply a force to a double-stranded RNA molecule while an RdRp transcribes the RNA. As progression of the fork converts double-stranded to single-stranded RNA in the tether, transcription can be followed by monitoring the vertical position of the bead. Figure panel adapted from [32**]. (d) Three traces of the transcription activity of an RdRp (blue points) acquired during the same experiment at an applied force of 20 pN and a NTP concentration of the order of mM, filtered with a 0.5 Hz low pass filter (black).
level (Figure 2c). Different traces taken during the same experiment differ remarkably (Figure 2d), and to characterize rare stable pauses large data sets were recorded at many experimental conditions. To analyze the data, a new approach based on the direct fitting of mechano-chemical models to complete data sets was used. We conclude by illustrating the basic ingredients of this approach, and comment on the signatures of the polymerase pause types found in the literature.

**Dwell-time based analysis for untangling molecular motor translocation pathways**

Standard approaches used to analyze single-molecule traces require either extensive pre-processing of traces in the form of pause-picking algorithms [3,17**], or discards important information about pause durations through considering local velocities [5,33]. Here we use simulations to outline a recent extension of dwell-time analysis that can be used to fit complete data sets directly to mechano-chemical models of translocating motors [32**]. In this approach, the traces are not segmented with fixed time intervals as is done in velocity-based analysis methods, but is segmented with fixed translocation distances — the dwell-time windows — to produce dwell-time statistics. As one often lacks the resolution to gather dwell-time statistics for single translocation steps, the data is instead collected over dwell-time windows that can span multiple translocation steps [47]. We will here focus on pause analysis, and not discuss the delicate and largely unresolved influence of noise on the shorter time events, such as pause-free translocation. For illustrative purposes, we will assume that on the time scale of pauses it is possible to smooth the trace sufficiently to suppress the effect of noise.

In Figure 3a we show a simulated example trace with a dwell-time window of 10 bp, and the series of scored dwell-times indicated as \( r_1, r_2 \), among others. Given the rates in any specific translocation scheme, one can in principle use first-passage time analysis [34] to calculate the DTD \( P_n(t) \) of times \( t \) for taking one step. The DTD for taking \( n \) steps is then the \( n \)-fold convolution \( P_n(t) = P_1^*P_1^*\ldots*P_1^* \) of the single-step distribution. The DTD can in principle be arbitrarily complicated, but we here focus on a few of the simplest important cases that result from translocation schemes suggested in the literature.

**First order process.** In the simplest scenario, each step along the elongation pathway is a first order process with a constant rate \( k \) (Figure 3b, inset). The one-step DTD is then an exponential distribution, and the \( n \) step distribution is a Gamma distribution of order \( n \) and characteristic rate \( k \). From Figure 3b it is clear that the width of the Gamma distribution depends on the number of substeps, a fact that can be used to estimate \( n \).

**Off-pathway pauses.** Polymerase elongation can be temporarily halted by thermally induced changes in enzyme structure [3,17,33], inducing what is referred to as off-pathway pauses. At its simplest, the structural change occurs with a first-order rate \( k_p \) and is reversed with the rate \( k_e \) (Figure 3c, inset). Though it is straightforward to calculate the DTD for arbitrary rates, again for simplicity, we focus on the case where pauses are entered infrequently in each dwell-time window, and typically last much longer than it takes to cross the dwell-time window without pausing. In this case, the first order process of the main elongation pathway captures most dwell-time windows, resulting in a Gamma distribution for short dwell times (Figure 3c). For longer dwell times, infrequent entrance into the pause state results in exponentially distributed dwell-times with a lifetime \( 1/k_p \). The probabilistic weight under the short-time Gamma distribution and the long-time exponential distribution reports on the probability to move through a dwell-time window without pausing and the probability to enter at least one pause. Importantly, this type of pause will not have a lifetime that depends on nucleotide concentration, as it is exited through the reversal of the thermal fluctuation that triggered the pause. The probability of entering a pause will depend on nucleotide concentration though, as the pause entry directly competes with nucleotide addition.

**On-pathway pauses.** Another type of pause observed in the literature results from a stochastic and drastic reduction of the nucleotide addition speed, for example, after a non-cognate base has been inserted [32**,35,36]. This leads to an on-pathway pause, which has a DTD similar to the off-pathway pause but with a pause-escape rate that is sensitive to nucleotide concentrations and an entrance probability that is independent on overall nucleotide concentration. Based on the different nucleotide dependencies of on-pathway and off-pathway pauses, it is possible to distinguish them using nucleotide concentrations sweeps.

**Hybrid pauses.** A recent study [32**] suggests that there exist off-pathway pauses, but where the paused state retains a degree of enzymatic activity. For such pauses, both pause probability and pause exit rate will depend on the overall nucleotide concentration.

**Backtracking pauses.** Another important pause is the backtracking pause, which results from the polymerase diffusively moving backwards and misaligning the product terminal with the active site. To reverse such a pause, the polymerase has to diffuse back to its original position [1,2]. The diffusive return induces pauses with a distinct \(-3/2\) power-law distributed dwell-times over a large time range [1,37*,38] (Figure 3d, inset). The width of this region is bounded by a lower corner time and upper cutoff time, both set by the hopping rates in the diffusive backtrack (Figure 3d) [37*].

**Composite distributions.** Experimental DTDs are generally more complicated than any of the above scenarios, both in
Dwell-time distributions for translocating systems. (a) Position versus time trace generated through a Gillespie simulation of a stochastic stepper with a simple pause (reaction scheme in inset of panel (c)). The times it takes to cross equal-sized dwell-time windows \((t_1, t_2, \ldots)\) are scored as dwell-times. (b) The distribution of dwell-times generated from the translocation scheme sketched in the inset is given by the Gamma distribution. Keeping the average time \(r\) it takes to cross the dwell-time window constant, the width of the distribution decreases with the number of steps needed to cross the window. (c) For a translocation scheme with a rare and long off-pathway pauses (reaction scheme in the inset), the dwell-time distribution will for short times follow a Gamma distribution centered on the average time it takes to translocate through a dwell-time window by \(n\) non-pausing steps \((n/k + k_0))\). At long dwell-times, the distribution will be dominated by the exponentially distributed pause-escape times (with the average \(1/k_0\)). The probabilistic weight of the Gamma and exponential-pause regions can be used to calculate the pause frequency. The basic shape of the dwell-time distribution of on-pathway pauses is the same, while its dependence on, for example, nucleotide concentrations differ (see main text). (d) For a translocation scheme including rare backtracks (reaction scheme in the inset), we again have an initial pause-free region that follows a Gamma distribution. At later times, there is an algebraically distributed region decaying with the power \(-3/2\), and bounded by the timescales \(t_0 = (k_0k_{out})^{-1/2}\) and \(t_1 = (k_{in}^{1/2} - k_{out}^{1/2})^{-2}\) [37]. The red dots in panel (c) and (d) are generated by Gillespie simulations of the translocation schemes shown in the insets in the respective panels.

that the timescales might not be well separated and that several types of pauses can be present at once. Still, the large data sets now becoming available through multiplexed experiments makes it possible to distinguish a large number of features (Figure 4). Details regarding the precise reaction scheme can be extracted not only by noting that pause mechanisms can be differentiated by how their probabilities and lifetimes depend on nucleotide concentration, but also from force dependence, and from the introduction of non-native nucleotide analogs [32**]. In Figure 4 we show the empirical DTD based on successively larger data sets collected from a viral RdRP [32**] with several pause types. An increasingly significant and rich structure pertaining to long-time and rare events is seen as the data sets are enlarged.

To fit models to DTD it is often convenient to determine the number of pause states through the Bayesian Information Criterion (BIC) [39], and fit out model parameters by Maximum Likelihood Estimation (MLE) [40] applied directly to the dwell-times. At present, this approach requires at least a partial analytic solution for the first-passage time of the corresponding reaction scheme [32**].
though future developments of numerical approaches will likely eliminate this requirement.

Outlook
The approaches discussed here can be applied to many different types of molecular motors. For example, replication is regularly interrupted by rare and slow events, such as proofreading activity on the nascent DNA and polymerase exchange. A recent study has also demonstrated the utility of combining high-resolution single-molecule techniques with next generation sequencing technology \[4^\circ\] to report on the specific DNA sequence responsible for a regulatory transcriptional pause. With further progress in multiplexing and the ability to localize the absolute position of a molecular machine on its template, it should soon become possible to investigate sequence-dependent pausing patterns at a global scale.

Conflict of interest statement
We have no conflicts of interest.

Acknowledgments
We thank our current and former colleagues in the Nynke Dekker and Martin Depken groups of the Department of Bionanoscience, TU Delft for useful discussions, in particular Jelmer Gnossen, Behrouz Esfandi-Mossallam, and Theo van Laar. We also acknowledge Elio Abbondanzieri for discussions on RNA polymerases, and Eric Suijler, Clara Posthumu, and Craig Cameron for discussions on RNA-dependent RNA polymerases. M.D. acknowledges early discussions with Stephan Grill and Eric Gallurt regarding dwell-time distributions and polymerases. This work was financed by a TU Delft startup grant to M.D. and by VICI and TOP grants from the Netherlands Organisation for Scientific Research to N.H.D.

References and recommended reading

- of special interest
- of special interest

5. First combination of single-molecule optical tweezers with new generation sequencing.


41. Huhle A, Klaus D, Brutzer H, Daldrop P, Joho, Otto O, Keyser UF, ** Seidel R: Camera-based three-dimensional real-time particle tracking at kHz rates and Angstrom accuracy. *Nat Commun* 2015, 6:5885. First high resolution magnetic tweezers, opening up the possibility of multiplexing data acquisition at high resolution.


