Engineering Dynamic Behavior into Nucleic Acids Guided by Single Molecule Fluorescence Microscopy

by

Jieming Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemistry)
in the University of Michigan
2019

Doctoral Committee:

Professor Nils G. Walter, Chair
Associate Professor Julie S. Biteen
Professor Robert T. Kennedy
Professor Kristen J. Verhey

Jieming Li

jmli@umich.edu

ORCID iD: 0000-0003-4536-4628

To my family and friends.

ACKNOWLEDGEMENTS

As a first-generation Doctoral student in my family, I feel very lucky and grateful for where I am now. Without the support and mentorship from many people along the way, I would not be able to arrive at this point.

I feel honored to have the opportunity to work with Dr. Nils Walter, who set a good example for his students of how much a great scholar could achieve. His high standard of scholarship and work ethic helped me establish my research skills and a scientist’s mindset. I want to specially thank him for the considerable time he has devoted and support he has provided to help me grow into an independent researcher.

The research I did is part of collaborations among research groups from Arizona State University (Hao Yan’s Lab) and Harvard University (William Shih’s Lab). Special thanks to Dr. Renee Yang for DNA tile fabrication support. This dissertation is a testament to their work as well as mine.

A good graduate student mentor is of large importance for people new to a lab. I have been truly lucky to have excellent graduate student mentors both in my undergraduate and graduate time. Thanks to Dr. Haixin Lin’s patience and guidance, I had the chance to learn what scientific research is really about during my undergraduate time. Another great person I owe many thanks to is Dr. Alex Johnson-Buck, who has been a patient and knowledgeable teacher when I joined the lab, a nice and helpful colleague during our collaboration, and a great mentor for both my research and life. I learned so many things from him.

Finally, I want to thank my family and my friends, without whom I would not have been able to get through all the hard moments on this trip. There have been times up and down on this journey. The excitement when things worked and the frustration when things didn’t helped me build skills necessary to be an independent researcher, and together formed an irreplaceable memory. I am very glad that my trip of exploring the fun in science becomes ever more delightful, and will never end.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Single-molecule fluorescence microscopy is a powerful technique that has been used for investigating the structural dynamics of biomolecules, and is particularly useful when ensemble averaging might obscure detailed information of the system under investigation. One application of single molecule measurement is to optimize the design of DNA nano-devices.

Dynamic DNA nanotechnology has yielded nontrivial autonomous behaviours such as stimulus-guided locomotion, computation, and programmable molecular assembly. Despite these successes, DNA-based nanomachines suffer from slow kinetics, requiring several minutes or more to carry out a handful of operations. In this thesis, I have pursued the speed limit of an important class of reactions in DNA nanotechnology—toehold exchange—through the single-molecule optimization of a novel class of DNA walker that undergoes cartwheeling movements over a field of complementary oligonucleotides. I identified the walking mechanism by single-molecule fluorescence resonance energy transfer (smFRET) measurement, with the stepping rate constant approaching 1 s^{-1}, which is 10- to 100-fold faster than prior DNA walkers. I also used single-particle tracking to demonstrate movement of the walker over hundreds of nanometers within 10 min, in quantitative agreement with predictions from the stepping kinetics. These results suggest that substantial improvements in the operating rates of broad classes of DNA nanomachines utilizing strand displacement are possible.

Another application of single molecule measurements is kinetic fingerprinting detection. Conventional methods for detecting small quantities of nucleic acids require amplification by the polymerase chain reaction (PCR), which necessitates prior purification and introduces copying errors. While amplification-free methods do not have these shortcomings, they are generally orders of magnitude less sensitive and specific than PCR-based methods. In this thesis, I review important experimental tips and data analysis details to provide a practical guide to a novel amplification-free method, single-molecule recognition through equilibrium Poisson sampling (SiMREPS), that provides both single-molecule sensitivity and single-base selectivity by monitoring the repetitive interactions of fluorescent probes with immobilized targets. In addition to demonstrating how this kinetic fingerprinting filters out background arising from the inevitable nonspecific binding of probes, yielding virtually zero background signal, I also investigated the detection of epigenetic mutations such as CpG methylation using SiMREPS.

The analysis of single-molecule microscopy data can be very time-consuming because there is no sufficiently robust automatic method for selection of qualified single-molecule fluorescence trajectories from the generally noisy and heterogeneous raw data, necessitating manual trace selection that can take hundreds of hours for large datasets. In this thesis, I discuss the innovative use of the popular convolutional neural

network AlexNet and the recurrent neural network Long Short-Term Memory (LSTM) to develop an automatic selector for single-molecule fluorescence resonance energy transfer (smFRET) traces. The average prediction accuracy is above 90% when tested on datasets from different users and experimental systems. To boost the selection accuracy and increase the diversity of training datasets, simulation data were included into the training data set and tested for selection accuracy. I expect that this new method will not only greatly expedite analysis of smFRET data and increase analysis reliability of SiMREPS data, but also introduce and validate machine learning as an effective tool for analysis of single-molecule microscopy data more generally.

Together, these results provide new insights into how single molecule microscopy can be used to engineer dynamic behaviors of nucleic acids.

Chapter 1

Progress towards Using Single Molecule Fluorescence Microscopy to Nanomachine Design and Diagnostic Detection

1.1 Introduction

Since the first observation of fluorescence in the 1500s, this natural wonder has inspired scientists to develop powerful techniques such as fluorescence spectroscopy and fluorescence microscopy. Among them, single molecule fluorescence microscopy has enabled scientists to observe the micro world with a closer look. For example, with assistance from single molecule fluorescence resonance energy transfer, molecular dynamic changes over a small distance (~7nm) can be measured. This feature can be used to help design and optimize complex DNA structures such as DNA tiles and DNA origami. Another application of single molecule microscopy is to interpret temporal fluorescence intensity changes of single molecule for diagnostic purposes, for instance, during the detection of cancer biomarkers such as microRNAs and DNA mutations.

When conducting such experiments, data analysis is very critical. Traditional single-molecule data analysis is time consuming and sometimes not satisfying the desired selectivity parameters. In addition, the inconsistency of raw data screening between operators can introduce subjective biases. As a recently thriving data analysis technique, machine learning is well known for its capacity for handling large datasets of high complexity, which provides a new option for single molecule fluorescence data analysis. In this chapter, I introduce single molecule fluorescence microscopy, DNA nanotechnology, nucleic acid detection using transient binding events, and the current data analysis in the single molecule fluorescence field as background to my dissertation.

1.2 Single Molecule Fluorescence Microscopy

Before the 2014 Nobel Prize in chemistry drew people's attention to super-resolved fluorescence microscopy, there had been remarkable growth in the use of single molecule fluorescence spectroscopy for the past decades. In fact, single molecule microscopy has begun to revolutionize the way people learn about micro biology systems. Methods such as reversible saturable optical fluorescence transitions (RESOLFT), photo activation localization microscopy (PALM) and point accumulation for imaging in nanoscale topography (PAINT) have broadened the power of optical microscopy by achieving a spatial resolution of approximately 20-30 nm[1–5]. Compared to traditional bulk ensemble assays to study complex biological systems, which usually read out the averaged information of the population of molecules, such a fine spatial resolution can reveal the heterogeneous behaviors among different molecules. This unique capability makes single molecule microscopy an ideal method to study dynamic behaviors of nucleic acids.

Fluorescence Resonance Energy Transfer (FRET) is a phenomenon by which non-radiative energy transfer occurs between two fluorophore molecules, a donor and an acceptor. The efficiency of FRET is extremely sensitive to small distance changes, being inversely proportional to the 6th power of the distance between the donor and acceptor. Specially, the FRET efficiency is described by $E = (1 + (R / R_0)^6)^{-1}$, where R is the inter-dye distance, and R_0 is the Förster radius, defined as the distance at which $E = 0.5$.

Single Molecule FRET (smFRET) is one of the most widely used and versatile methods to study the features of individual biological molecules, especially involving asynchronous dynamic behaviors[6]. In particular, tracking the FRET efficiency in real-time can report on the conformational dynamics of single molecules. Single-molecule FRET data are usually acquired using a wide-field total internal reflection fluorescence (TIRF) microscope, which can achieve high-throughput data sampling compared to confocal microscopy[7,8]. One advantage of smFRET is its capability of ratiometric measurements of the internal distance on a molecular scale, largely reducing systematic errors that may result from instrument noise and drift[9,10]. Also the subnanometer structural kinetic information that smFRET provides is not usually available through any other methods.

Single molecule FRET is increasingly used in studies of DNA nanotechnology, a field whose goal is the rational design of DNA molecular structures and dynamic devices[11]. As any engineering discipline this effort requires detailed feedback of the building process and products, information that smFRET can often uniquely provide. The dynamic information acquired using smFRET reflects a distribution of properties (rather than an average) because the measurement is conducted on individual molecules. This feature of smFRET enables the study of tiny structures associated with dynamic DNA devices, reporting on diverse dynamic behaviors.

In my thesis, smFRET has been used as a tool to investigate the translocation mechanism of a new type of DNA walker; single particle tracking is used for lager area observation of the DNA walker; a kinetic fingerprinting detection method inspired by a super resolution imaging technique is further studied for method optimization and DNA methylation detection; analysis on data generated from single molecule measurement is studied. Overall, the single molecule microscopy is the driving force of my thesis research.

1.3 DNA Nanotechnology and Nanodevices Design

The field of DNA nanotechnology can be traced back to 1980s when people wanted to organize proteins in 3D crystals by using DNA as the connecting bones[12]. Since then, DNA nanotechnology has been developed as a reliable technique in control of matter on the nanoscale. Besides DNA self-assembly control and design as the major interest of researchers[13,14], using DNA to build artificial molecular motors is also an important application of DNA nanotechnology.

Since the step-by-step (hand-over-hand) movement mechanism of molecular motors such as dynein, myosin and kinesin super families was characterized[15,16], attempts to mimic their dynamic behaviors have been made in the form of synthetic molecular walkers. Several DNA-based molecular walkers have been synthesized[17–23], motivated by a long-term goal of controlling molecular transport processes with the programmability and structural robustness of DNA nanotechnology. Previous studies have shown that DNA walkers can walk directionally along a track upon sequential addition of a DNA strand as chemical "fuel"[19,24]. In some studies, more sophisticated tasks are coupled to walker motion, such as templating sequential chemical reactions and assembling gold nanoparticles[25,26].

Despite all this progress, the DNA walkers reported so far have been constrained by slow translocation rates, which are typically on the order of a few nm/min[17,23]. By comparison, natural protein motors have translocation rate of ~1μm/s under saturating ATP conditions, a 3~4 orders of magnitude faster rate[15,27]. It is desirable to reduce this gap if synthetic DNA walkers are to serve as useful agents of molecular transport.

Based on prior investigations, the translocation rate of many DNA walkers is believed to be limited by slow catalytic steps or the release of cleavage products. In contrast, the displacement of one strand DNA duplex by another can be catalyzed by the nucleation of short single-strand overhangs, or “toeholds”, a process that can be very rapid when the reagents are present at high concentrations[28]. This process is the so-called “toehold exchange displacement”. Since DNA scaffolds can be used to generate local effective reagent concentrations in excess of 100 μM in bimolecular reactions[29], a DNA walker using toehold exchange reactions for locomotion may be able to realize higher translocation rates than previously reported DNA walkers.

In my thesis, toehold exchange displacement is used as a new translocation mechanism to speed up the DNA walkers’ movement. The fast movement of DNA walkers is achieved by allowing the DNA walkers to move along the foothold strands with a cartwheeling movement fashion.

1.4 Hybridization patterns of nucleic acids: molecular kinetic fingerprints

The detection of nucleic acid sequences with high specificity plays an important role in both basic biological research and diagnostics due to the fundamental roles of genetics, epigenetics, and gene expression in both normal physiology and pathology. For instance, specific mutations and aberrant methylation patterns of DNA have been linked to various types of cancer, showing promise for early detection of disease, monitoring of treatment response and relapse, and indicating whether a cancer is likely to respond to a given course of treatment. Expression levels of specific microRNAs (miRNAs) and long non-coding RNAs (lncRNAs) are strongly correlated to cell differentiation states and thus of interest as biomarkers of disease. MicroRNAs (miRNAs) are a class of small non-coding RNA molecules found in plants, animals, and some viruses. They functions in RNA silencing and post-transcriptional regulation of gene expression.

To ensure adequate sensitivity for most nucleic acid analyses, samples must be amplified by procedures such as the polymerase chain reaction (PCR) for adequate specificity, sometimes following generation of cDNA by reverse transcriptase and/or the ligation of adapter sequences. However, such preparative procedures introduce several challenges for the quantitative analysis of nucleic acids. First, DNA polymerases and thermal cycling can both introduce artifactual sequence changes such as base substitutions during the amplification process, which may result in false positives when attempting to detect rare single-base mutations (e.g., for liquid biopsy of cancer). Second, reverse transcriptases and ligases exhibit significant sequence biases, introducing significant artifacts such as spurious differences in expression levels and even the complete absence of certain sequences. Third, polymerases and ligases can be susceptible to inhibition by contaminants such as heparin and heme, necessitating additional purification steps prior to amplification. Finally, many classes of analytes are simply not amenable to direct amplification, including epigenetic modifications, short or fragmented nucleic acids, or non-nucleic acid analytes.

Conventional methods for detecting small quantities of nucleic acids require amplification by the polymerase chain reaction (PCR), which requires prior purification and introduces copying errors. While amplification-free methods do not have these shortcomings, they are generally orders of magnitude less sensitive and specific than PCR-based methods. To overcome these drawbacks, Johnson-Buck *et al*. developed a miRNA detection method based on kinetic fingerprinting called Single-Molecule Recognition by Equilibrium Poisson Sampling (SiMREPS). By monitoring transient binding of fluorescent probes to immobilized miRNA targets using TIRF microscopy, the authors demonstrate an amplification-free, zero-background, single-molecule detection and counting technique. In contrast to other amplification-free nucleic acid detection methods, whose specificity is limited by thermodynamic discrimination between true and spurious targets, this SiMREPS technique largely overcomes the inherent thermodynamic barriers through repeated kinetic sampling. Johnson-Buck *et al.* performed SiMREPS detection of five exemplary miRNAs and realized a more than 500-fold discrimination between two members of the *let-7* miRNA family differing by a single nucleotide, *let-7a* and *let-7c*. They further demonstrated the rapid and specific detection of endogenous *let-7* in crude human cancer cell lysate and quantification of the clinically relevant *miR-141* in a matrix of minimally treated human serum[30].

Results from SiMREPS detection on miRNAs provides a firm foundation for further study and optimization. A 1 to 1 million specificity of wild type and mutant targets has been achieved from the detection of circulating tumor DNA (ctDNA) using SiMREPS[31]. Besides expanding the detectable analytes spectrum to DNAs or other targets, an important study direction is the measurement optimization, both in experiments and data analysis. In the experimental side, improving the mass transfer of analytes from low concentration samples and modification of imaging surface for different types of analytes are of high priority. In the data analysis side, finding a new way to achieve a higher selectivity among similar-look traces is critical.

In my thesis, optimization of both experiment and data analysis is explored. In addition, early investigation of detecting smaller scale difference (methyltation of DNA CpG site) among nucleic acid molecules is also tried.

1.5 Time series data analysis in single molecule microscopy

The output of single molecule fluorescence microscopy measurement is usually displayed as a time series containing information regarding the dynamics and kinetics of the system as intensity vs. time format traces. In daily research, a manual screening of traces is always needed to get a good quality dataset before the data will be used in further analysis. Sorting and classification of single-molecule time series data (such as single-molecule fluorescence resonance energy transfer data) is a critical but time-consuming part of analyzing single-molecule measurements. This is because the large diversity of potential background signals makes it difficult to design simple criteria, such as thresholds based on intensity or noise, that would effectively remove all irrelevant time traces (e.g., those resulting from contaminants or nonspecific binding rather than analyte molecules) while retaining all or most of the relevant data for further analysis. Furthermore, it is often the case that only a particular time segment of each single-molecule trace is useful, and these regions of interest (ROIs) must typically be selected by hand, which slows down analysis considerably. Because the selection of ROIs is dependent on relatively complicated determinations – such as the absence of various artifacts including photobleaching, blinking, and contamination from nearby fluorescent materials – it is also difficult to automate this process using conventional methods. In short, the time series data analysis discussed here can be addressed as a sorting and segmentation task.

In my thesis, the obstacles of sorting of multi-channel data (more than one signal in the time series data such as data from smFRET measurement) and single channel data generated from single molecule measurement, and the solutions using machine learning algorithms developed together with my collaborators are discussed.

1.6 Machine learning and its general applications

As a currently highly popular data analysis method, machine learning has been widely used in science, economics, business and other areas[32–40]. Machine learning systems are used to identify objects in images, transcribe speech into text, match news items, posts or products with users' interests and select relevant results of search. Traditional machine learning requires careful engineering and considerable expertise to design a feature extractor, which can convert the raw data into suitable representations or features that can be easily detected or classified by the learning systems. Deep learning is an advanced format of machine learning in the way that it allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Because of the unique non-linear modules (neural network layers) that can extract features at a higher and more abstract level, complex functions can be learned in deep learning neural networks. The key difference of deep learning from traditional machine learning is that the layers of features are not designed or engineered by human beings, but are learned from data using a general-purpose learning procedure. This difference makes deep learning a preferable option for researchers outside the statistics field and more applicable to diverse tasks. Besides the beating traditional machine learning in traditional fields such as image recognition and speech recognition[41–46], deep learning has also beaten other machine learning algorithms at scientific research tasks such as predicting the effects of mutations in DNA on gene expression and disease[47,48].

The similarity of sorting traces and machine learning clustering, and the reality of no automated process to handle the sorting task, make single molecule trace selection an exciting opportunity to use machine learning. In our case, recurrent neural network (RNN) and Convolutional neural network (CNN) are being used. Both networks are deep learning networks. But since we are using these algorithms to solve a data analysis problem rather than developing a new algorithm, and to better reach a broader audience, we will use the general term machine learning in this dissertation even though sometimes we are referring to deep learning.

1.7 Overview of the dissertation

In the following chapters, I will discuss several applications of single molecule fluorescence microscopy such as to study individual nanoscale devices composed of DNA and to develop diagnostic detection method.

In Chapter 2, I describe the mechanistic study of a new type of autonomous DNA walker that uses toehold exchange displacement. By using single molecule fluorescence resonance energy transfer (smFRET) measurement, we are able to optimize the design to get the fastest DNA walker. We also demonstrate that this DNA walker can translocate in a much larger two dimensional area in a fast speed by single particle tracking observation.

In Chapter 3, I will discuss how binding kinetic information can be used for diagnostic detection through single molecule measurement. By recognizing the specific binding kinetic patterns of different analytes, this kinetic fingerprinting method has achieved a 1 to 100,000 selectivity. This method is discussed thoroughly from experimental execution to data analysis details. A practical guide of using this method is embedded in this chapter. One new application of this kinetic fingerprinting method to directly detect DNA CpG methylation is also discussed.

In Chapter 4, a collaborator and I have developed an automatic single molecule traces selector by the innovative use of machine learning algorithms. Two types of traces are studied here: single channel traces from kinetic fingerprinting measurement and double channel traces from smFRET measurements. In analyzing single channel traces, we show that the new machine learning method has the potential to outperform the current kinetic thresholding, which uses Hidden Markov Modeling (HMM). In double channel trace analysis of smFRET measurements, the time spent on manual trace selection can be considerably reduced, together with an increase of consistency and objectivity in raw data screening. Overall, productivity and specificity in single molecule data analysis can be improved by this new trace selector.

Together, the work presented in this dissertation contributes to the 1) analytical toolkit of DNA nanotechnology; 2) diagnostic detection method; and 3) single molecule data analysis. The main aim of my thesis is to apply single molecule measurement to different uses such as a new characterization method for mechanistic study and optimization of a DNA walker, and developing a new direct detection method for epigenetic mutation. With improvement from data analysis, single molecule fluorescence microscopy promises to become an even better tool to engineer dynamic behaviors of nucleic acids.

Chapter 2:

DNA Acrobat as A New Tool to Explore the Speed Limit of DNA Walkers[1,2]

2.1 Introduction

Dynamic DNA nanotechnology exploits the programmable reconfiguration of Watson-Crick base pairing to carry out nontrivial autonomous functions inspired by both biology and macroscopic engineering, resulting in systems such as molecular walkers[24,49,22,50,51], assembly lines[52,53], computers[54,55], and robots[17,56,57]. Unlike their naturally occurring counterparts, the top-down design and transparent relationship between sequence and function of DNA nanodevices provides rich opportunities for programmable specificity and dynamics. A fundamental process in nearly all such systems is strand displacement, the stepwise replacement of one strand of a double helix with another invading strand, a process often catalyzed by one or more short overhangs of unpaired nucleotides called toeholds[58]. A class of strand displacement reactions called toehold exchange[28] involves competition between two toeholds of similar length, and has found widespread use in dynamic DNA nanotechnology[54,57,59] due to its robust ability to

[1] Reproduced in part from Li, J., Johnson-Buck, A., Yang, Y. R., Shih, W. M., Yan, H., & Walter, N. G. Nature Nanotechnology 13, 723–729. Copyright Nature Publishing Group, 2018.

[2] Renee Yuhe Yang designed, fabricated, and characterized DNA tile samples. William M. Shih. and Alexander Johnson-Buck designed, fabricated, and characterized DNA origami samples. Jieming Li performed smFRET and single-particle tracking measurements. Alexander Johnson-Buck performed kinetic Monte Carlo modeling. Jieming Li, Alexander Johnson-Buck, and Nils G. Walter analyzed and interpreted the data. Jieming Li and Alexander Johnson-Buck co-wrote the paper.

accelerate the exchange of nearly isoenergetic DNA double helices by several orders of magnitude[28]. Indeed, toehold-mediated strand displacement reactions can have second-order rate constants in excess of 10^6 $M^{-1}s^{-1}$; thus, DNA nanomachines with local effective strand concentrations in the micromolar range[29] may be able to execute individual operations in seconds or less, provided that branch migration and toehold dissociation are sufficiently rapid.

Despite this theoretical rapidity, in current practice most DNA nanomachines require several seconds to several hours to complete a single operation[52–54,60]. For instance, a recently reported cargo-sorting DNA robot utilizing toehold exchange for locomotion was found to take about one step every 5 min despite an only 6-nm gap between neighboring footholds[57], which is comparable to the reported speeds of other autonomous DNA walkers[17,18,60]. In one notable exception, a translocation rate of ~1 µm min^{-1} was achieved for DNA-functionalized nanoparticles, though this utilized a burnt-bridge mechanism involving the degradation of complementary RNA strands by RNase H[61]. By comparison, natural protein motors have translocation rates of ~1 µm/s under saturating ATP conditions[27,15]. We hypothesized that the sluggish performance of DNA nanomachines is not due to a fundamental limitation of strand displacement reactions, but is instead the result of designs not optimized for speed.

To test this hypothesis, we designed a novel type of DNA walker with the express purpose of rapid locomotion. We then used single-molecule fluorescence resonance energy transfer (smFRET) to characterize its translocation mechanism and kinetics, and optimized its stepping rate by systematically varying the lengths of its toehold and branch-migration domains, resulting in stepping rate constants more than an order of magnitude faster than existing DNA walkers. Following optimization of the design with smFRET, we used single-particle tracking to observe the movement of toehold exchange DNA walkers over distances as long as ~1 μm on a two-dimensional array of footholds. The performance of the walkers on 2D arrays is quantitatively consistent with predictions based on the stepping kinetics measured by smFRET on DNA tile substrates, demonstrating that this mechanism of locomotion is generalizable to different substrate types and conducive to long-range movement requiring hundreds of steps.

2.2 Materials and Methods

2.2.1 Materials

Single-stranded oligonucleotides, Cy3- and Cy5-labeled oligonucleotides, and amine-modified oligonucleotides were purchased from IDT (Integrated DNA Technologies, Inc.). Dye-labeled oligonucleotides were HPLC-purified by the manufacturer. Streptavidin (S-888), biotinylated bovine serum albumin (bBSA, 29130), Trolox (218940050), and 3,4-dihydroxybenzoic acid (AC114891000) were purchased from Thermo Fisher Scientific. Protocatechuate 3,4-dioxygenase (PCD), tris base, acetic acid, EDTA, magnesium acetate, dibenzocyclooctyne-N-hydroxysuccinimidyl ester (761524), 11-azidoundecyltriethoxysilane (SIK4711-30), *N,N*-dimethylformamide (DMF), triethylamine (TEA), and sodium acetate (NaOAc) were purchased from Sigma-Aldrich.

2.2.2 Design, Assembly, and Characterization of 4-Helix DNA Tiles

The detailed sequence designs of the DNA 4-helix (4HX) tile nanostructures are shown in Figures A1.1, A1.2. The computer program Tiamat was used for structural design and sequence generation. Oligonucleotides were purified using 6-8% denaturing polyacrylamide gel electrophoresis (PAGE) at room temperature. The bands corresponding to the correct strand length were imaged by UV lamp (254 nm) and then cut from the gel, chopped into small pieces, and incubated overnight in elution buffer (500 mM ammonium acetate, 10 mM magnesium acetate, 2 mM sodium ethylenediaminetetraacetic acid, pH 8.0). The DNA strands were extracted from the gel pieces by centrifugation using a Costar Spin X filtration device (Corning, cellulose acetate membrane with 0.22 μm size). The filtrate was then ethanol-precipitated, washed by ethanol and dried under vacuum. The DNA strands were dissolved in nanopure water and the concentrations of the individual purified strands were measured by UV absorbance at 260 nm using the extinction coefficient provided by the manufacturer.

The DNA strands constituting each DNA structure were mixed in 1×TAE/Mg^{2+} buffer (40 mM Tris, 20 mM acetic acid, 2 mM EDTA and 12.5 mM magnesium acetate, pH 8.0) to reach a final concentration of 1 μM per strand. All samples were annealed using an Eppendorf Mastercycler using the following annealing protocol: heat to 90 °C, cool from 90 °C to 72 °C over 10 min, then from 68 °C to 24 °C over 60 min, and finally holding at 15 °C.

The formation of the DNA structures was characterized by native PAGE. 5% Native PAGE gels were prepared at room temperature and run for 4 to 6 hours at a constant voltage of 200V. Cy3-labeled structures were visualized by UV lamp (365 nm) and then cut from the gel, chopped into small pieces, and incubated overnight in 1×TAE/Mg^{2+} buffer. The DNA structures were then extracted from the gel pieces by centrifugation using a Costar Spin X filtration device (Corning, cellulose acetate membrane with 0.22 µm size). For analytical native PAGE, the gel was subsequently stained with SYBR® Green/Gold.

2.2.3 Design, Fabrication, and Characterization of DNA Origami

The DNA origami structure was built on a square helical lattice using the program caDNAno, and staple strand sequences (Table A1.1) were designed to be complementary to a contiguous segment of the M13 p7308 scaffold sequence. Assembly was performed by combining 10 nM of the p7308 scaffold with a tenfold molar excess of all staple sequences in TE buffer with 10 mM $MgCl_2$, then performing the following annealing protocol on a Tetrad 2 Peltier thermal cycler: heat to 80 °C, decrease to 60 °C over 70 min, decrease from 60 °C to 24 °C over 66 h, and then hold at 4 °C. Origami were purified from excess staples by 2% agarose gel electrophoresis in 0.5X TBE + 10 mM $MgCl_2$. The gel was scanned using a Typhoon FLA 9000 (GE Healthcare Life Sciences), and a scale printout was laid under the gel to permit excision of the origami bands. Origami structures were eluted from the gel by centrifugation for 5 min at 5000 × g in Freeze ‘N Squeeze spin columns (Bio-Rad). Recovery was confirmed by again running the purified origami on a 2% agarose gel and scanning for Cy3 fluorescence. DNA origami morphology was characterized by negative-stain transmission electron microscopy (TEM) using a JEOL 1400 TEM after depositing 3 µL of origami on a plasma-treated carbon Formvar grid (Electron Microscopy Sciences) and staining with a freshly prepared 2% uranyl formate solution for 0.5 min.

2.2.4 smFRET characterization of DNA walkers on DNA nanostructures

Microscope slides with a flow channel were prepared using double-sided tape (3M) and treated with biotinylated BSA and streptavidin as described[62,63] to prepare the surface for immobilization of biotinylated DNA nanostructures. 15 μL of a 10 nM solution of DNA tile or origami and 10 μL of a 10 nM solution of DNA walker were combined and incubated in dark at 37 °C for 5 minutes. This mixed sample was diluted to 20 pM in TA-Mg^{2+} (40 mM Tris, 20 mM acetic acid, and 12.5 mM magnesium acetate, pH 8.0) buffer, then injected into the flow chamber. After incubating for 10 minutes, TA-Mg^{2+} was injected to remove excess unbound material.

Single-molecule FRET experiments were carried out on an inverted prism-type total internal reflection fluorescence (TIRF) microscope with a 1.2 NA 60× water-immersion objective (IX71, Olympus) in a darkened room at an environmentally controlled temperature of 20 ± 3 °C. Fluorescence excitation was provided by a 532-nm green laser (CrystaLaser CL532-050-L, 50 mW, attenuated and focused to give an illumination intensity of ~100 W cm^{-2} in the sample plane); presence of an active FRET acceptor was confirmed at the beginning of each experiment by brief excitation with a 640-nm red laser (Coherent CUBE 635-25C, 25 mW). The Cy3 and Cy5 emission signals were separated by a dichroic mirror with a cutoff wavelength of 610 nm (Chroma) and projected side-by-side onto an ICCD camera chip (iPentamax HQ Gen III, Roper Scientific, Inc.) with a full-frame acquisition rate of 10 Hz. The Cy3 channel image was passed through a bandpass filter (HQ580/60m, Chroma) and the Cy5 channel was passed through a long-pass filter (HQ655LP, Chroma). A Newport ST-UT2 vibration isolation table was used in all experiments to reduce instrument interference. In all smFRET measurements, an oxygen scavenger system (OSS ≡ 5 mM 3,4-dihydroxybenzoic acid; 2 mM Trolox; and 50 nM protocatechuate dioxygenase) was included in the imaging buffer to retard photobleaching.[29,64]

Analysis of single molecule FRET trajectories was performed with custom-written MATLAB scripts as previously described[65], with the FRET ratio at each time point calculated as $I_{Cy5}/(I_{Cy5} + I_{Cy3})$, where I_{Cy5} and I_{Cy3} are the apparent fluorescent intensities of Cy5 and Cy5, respectively. A given smFRET trajectory was used in subsequent analysis only if it (1) exhibited total fluorescence of Cy3 and Cy5 exceeding 500 counts/frame; (2) showed clear evidence of both Cy3 + Cy5; and (3) showed no evidence of multiple identical fluorophores, for example, multiple photobleaching steps or overlapping point-spread functions in the CCD image. After trajectories met the criteria were selected, Hidden Markov modeling (HMM) was then applied using the QuB software suite (State University of New York at Buffalo) to determine the mean dwell times in high- and low-FRET states for each trajectory.[66] The same two-state model was applied to all datasets. After idealization, the dwell times in the high- and low-FRET states (red and blue lines in Figure 2.2 f-i) were extracted from each trace, and the mean value of all high-FRET, low-FRET, or (high-FRET + low-FRET) dwell times were used to describe the mean dwell time of each molecule. These mean dwell times across all observed molecules are represented as box-and-whisker plots in Figure. 2.3 a,b. Transition occupancy density plots (TODPs), which depict the frequency of transitions from an initial FRET state to a different (‘final’) FRET state among all molecules characterized, were constructed from the idealized data as described previously[65].

Cross-correlation between donor and acceptor fluorescence signal (Figure 2.1e) was calculated using the built-in MATLAB function xcorr with unbiased normalization; the cross-correlation signal was further normalized such that the cross-correlation approaches 0 at infinite time lag, and any positive or negative correlation is confined to the interval [-1,1]. The decay time of cross-correlation was estimated by a single exponential fit.

2.2.5 Preparation and characterization of high-density Foothold-functionalized surfaces

22×50 mm coverslips were sonicated in a solution of 2% Alconox for 5 min and then rinsed 5 times with deionized water. Rinsed coverslips were incubated in heated base piranha solution (5% hydrogen peroxide and 5% ammonium hydroxide, 60-70 °C) for 40 min. Coverslips were rinsed 5 times with diH_2O following once with ethanol, and dried under an air stream. Dry coverslips were placed into a box with ethanol-soaked Kimwipes (Kimberly Clark). A 2% silane solution was prepared by combining 2 μL 11-azidoundecyltriethoxysilane and 98 μL ethanolic acetic acid (95% ethanol/5% aqueous acetic acid), and 80 μL of the solution was added to the coverslip. Another coverslip was placed on top to form a sandwich. After a 10-min incubation, the coverslip sandwiches were flipped over and again incubated for 10 min. After a total 20-min incubation, the coverslips were rinsed with absolute ethanol twice and dried under air.

Amine-modified oligonucleotides (***$F_{1,covalent}$***: 5′-CAATACCCCTACGGTCACTTCTTTTTTTTTT/3AmMO/; and ***$F_{1,covalent}$***: 5′-/5AmMC6/TTTTTTTTTTCCCTCATTCAATACCCCTACG)**,** were functionalized with DBCO as follows: 200 μL of *N,N*-dimethylformamide (DMF) and 5 mg of DBCO-NHS ester were combined to prepare a 62.5 mM solution of DBCO-NHS ester. A 20 µL aqueous solution of 1 mM amine-modified Foothold strand ***F₁*** or ***F₂*** was combined with 40 µL of 62.5 mM DBCO-NHS ester, 39 µL of DMF, and 1 µL of TEA and incubated for 2 h at room temperature. DNA was precipitated by adding 10 µL 3 M NaOAc and 200 µL of chilled 100% ethanol to the reaction, and incubating at -20°C for 30 min. The precipitated DNA was pelleted by centrifugation at 20,000 × g, 4°C, for 40 min. The supernatant was removed, and the precipitate was rinsed with 100 µL 80% ethanol and spun down again for 1 minute. The supernatant was again removed, and the pellet was dried in vacufuge for 10 minutes. A UV spectrum was collected to measure the concentration of DBCO functionalized oligos. The absorbance of DBCO at 310 nm was used to estimate the ratio of DBCO to DNA (approximately 1:1) after subtracting the absorbance of the DNA at 310 nm based on its absorbance at 260 nm and the extinction coefficients at 260 and 310 nm predicted using the UV Spectrum application of IDT biophysics.

Click conjugation of oligos to azide-functionalized coverslips was performed as follows. Solutions of 50 µM DBCO-functionalized Foothold oligonucleotides ***F_1*** and ***F_2*** were prepared in PBST buffer (1× PBS + 0.1% Tween-20). Equal volumes of each 50 µM oligonucleotide solution were combined, and 1 µL of the mixture was spotted onto the 11-azidoundecyltriethoxysilane-modified coverslips. The click reaction proceeded overnight for more than 15 hours in a humid environment. Coverslips were rinsed thoroughly by ddH_2O for more than 10 seconds and dried under N_2.

The density of click-conjugated oligonucleotides on coverslip surfaces was estimated as follows. For purposes of density characterization, only DBCO-modified ***F_1*** strands were added (at 50µM) during the click conjugation. After construction of sample wells (see Single-particle tracking and data analysis, below), 100 µL of a mixture containing 1 pM Cy5-labeled $\boldsymbol{W_{8\_13\_8}}$ and 1 µM (*i.e.*, a 1-million-fold excess) of a non-fluorescent $\boldsymbol{W_{8\_13\_8}}$ in TA-Mg^{2+} buffer was added to the sample well and incubated for 10 min. The solution was replaced by 100 µL of OSS and the sample was imaged on the same TIRF microscope used for single-particle tracking, under illumination at 640 nm. The number of Cy5-labeled walker molecules bound within each field of view was estimated using a custom MATLAB script, and averaged over 23 fields of view, and multiplied by 10^6 to estimate the total number of labeled and unlabeled $\boldsymbol{W_{8\_13\_8}}$ molecules per field of view, $N = 3.0\times10^8$. For each field of view, the area is S = (262nm/pixel × 512 pixels)^2 = 17990 μm^2. The density of Foothold oligos is N/S = 1.67×10^4 /μm^2.

To estimate the typical distance to the nearest available Foothold of the opposite type (since a walker with a dissociated toehold domain can step only onto F_1 if it is bound to F_2*, and vice-versa*), we performed numerical simulations of the distribution of footholds $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ in two dimensions (Figure A1.8a), assuming an independent uniform random distribution of each foothold on the surface and the same density of footholds as we observed experimentally (1.67x10^4 μm^{-2}; in the simulation, this translates to 8350 copies of F_1 and 8350 copies of F_2 in a 1000 nm x 1000 nm region). We then calculated the distance to the nearest F_2 for each foothold F_1, and plotted a histogram of these distances (Figure A1.8b). This simulation predicts that the mean distance to the nearest foothold is 5.5 nm, and >98.5 % of footholds will have at least one foothold of the opposite type within 13 nm.

2.2.6 Estimation of mean step size in 2D walking experiments

To estimate the mean step size, we approximated the mean distance between the foothold's anchor point and the distal (free) toehold of the walker as (length of walker-foothold duplex) + (length of single-stranded linker between coverslip and walker). The length of the walker-foothold duplex was calculated as (number of nucleotides) x (0.33 nm per nucleotide). The RMS end-to-end distance of the single-stranded linker (dT_{10} + any unpaired toehold nucleotides in the proximal foothold) was estimated using a freely jointed chain model according to the formula $RMSD = \sqrt{N}l$, where N is the number of Kuhn segments and l is the persistence length[67], assuming a persistence length of 1.5 nm and a contour length of 0.56 nm per nucleotide[68]. The estimated step sizes of the two walkers used in single-particle tracking are estimated as 10.8 nm (for $W_{6\_13\_6}$) and 11.1 nm (for $W_{8\_13\_8}$), which were used in generating the simulated MSD vs. time traces for Figure 2.4f. However, given the predicted contour length of the 10-12 nucleotide ssDNA linker (5.6-6.7 nm) and the predicted length of the 19-21 nucleotide dsDNA segment (6.3-7 nm), the maximum end-to-end distance of each walker is as large as 12.5-13 nm.

2.2.7 Single-particle tracking and data analysis

A 200-µL Eppendorf micropipet tip was cut with a razor blade and attached to the $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ modified coverslip by Epoxy (Double Bubble, Hardman Adhesives) as described[30] to form a sample chamber with the DNA-coated region positioned approximately in the in the center of the chamber. The sample chamber was incubated with 100 µL PBST buffer (1x PBS + 0.1% Tween-20) for 10 min, then for 15 min with a 100 µL mixture containing 1 pM DNA walker and 10 pM Cy3-labeled fiducial marker oligo (sequence fully complementary to $\boldsymbol{F_1}$) in PBST. The walker sample was then removed and the chamber was rinsed 3 times before imaging.

Single-particle tracking experiments were performed on an Olympus IX-81 objective-type TIRF microscope equipped with a 60× oil-immersion objective (APON 60×OTIRF, 1.49 NA) with both Cell^TIRF and z-drift control (ZDC2) modules, and an EMCCD camera (IXon 897, Andor, EM gain 300). Cy5 excitation was provided by a 640-nm red laser (Coherent CUBE 640-100C, 100 mW) and Cy3 excitation was provided by a 532-nm green laser (CrystaLaser CL532-150-L, 150 mW). In all single-particle tracking experiments, an OSS was included in the imaging buffer to retard photobleaching. The translocation of DNA walkers was monitored under alternating TIRF excitation at 640 and 532 nm (time lapse interval = 30 s, exposure time = 100 ms) for 60 min.

Analysis of single-particle tracking experiments was performed as follows. The ImageJ plug-in Particle Track and Analysis (PTA) was used to conduct 2-dimensional Gaussian fitting by the Levenberg-Marquardt method to obtain trajectories for each detected walker molecule.[69] The search area was set to 3 pixels (= 402 nm). The net movement of all fiducial markers in the field of view was subtracted from walker trajectories using a custom Matlab script to account for x-y stage drift. A given trajectory was used in subsequent analysis only if it (1) lasted 10 minutes (20 frames) without photobleaching; (2) exhibited no sudden fluorescence intensity changes as determined by manual inspection of the output from PTA fitting; and (3) showed no evidence of multiple identical fluorophores, such as multiple photobleaching steps or overlapping point-spread functions in the CCD image.

Calculation of mean square displacement (MSD) was performed as follows. An initial position (x_0, y_0) was defined as the arithmetic mean of the first 3 position measurements of each trajectory. The distance of each subsequent position measurement (x_i, y_i) from initial position (x_0, y_0) was then calculated and squared to obtain the squared net displacement over time. The arithmetic mean of the squared displacement (MSD) was calculated for all trajectories lasting at least 10 min, and the corresponding standard error (SE) of the mean was calculated for each MSD value as plotted in Figure 2.4f. After fitting a linear function to each MSD *versus* time plot in OriginPro 8.0, the slope of the linear fit was divided by 4 to obtain the apparent 2D diffusion coefficient (*D*) for each walker, based on the 2D diffusion model $\langle (x(t))^2 \rangle = 4Dt$. The diffusion coefficient was also predicted from smFRET measurements of stepping kinetics using the 2D random walk model $\Delta x^2 = 4D\Delta t$, in which Δx is the step size (assumed to be 10.9 nm) and Δt is the mean stepping lifetime of a representative (median-valued) walker.

2.2.8 Kinetic Monte Carlo modeling of branch migration and stepping kinetics in 3-foothold system

Branch migration and stepping of walkers in a 3-Foothold system was numerically simulated at single-base resolution using a version of the Gillespie algorithm[70] implemented in MATLAB. Additional details regarding the simulations and their interpretation are provided in the published paper. Autocorrelation of the branch migration state was calculated as a function of time lag using the xcorr function in MATLAB, and normalized in the same manner as for cross-correlation in the smFRET data.

2.3 Results

2.3.1 smFRET study of DNA walker's stepping mechanism on DNA tile

The DNA walker (Figure. 2.1a) was designed with speed, simplicity, and robust performance as the primary objectives. Thus, we intentionally avoided the need for strand cleavage by protein or DNA enzymes, since these pose the risk of creating kinetic traps for DNA walkers[71] and would prevent observation of repeated stepping. Instead, we chose a mechanism based purely on toehold exchange, which permits the walker to step an indefinite number of times between two competing "foothold" DNA sequences, undergoing rapid sequence-guided movement over long distances while remaining stably bound to at least one foothold at all times. Furthermore, we chose a "cartwheeling" mode of locomotion so that the free toehold is always distally located and can pivot about its anchor point to position its free toehold near an unoccupied complement for rapid binding, thus yielding comparable reaction rates for each step. As depicted in Figure. 2.1a, the single-stranded DNA walker $\boldsymbol{W}$ undergoes head-over-heels movement over a surface of two different foothold sequences, $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$, by toehold-mediated strand displacement. The two foothold sequences comprise a common middle branch migration domain $\overline{\boldsymbol{D_B}}$ as well as toehold domains $\overline{\boldsymbol{D_A}}$ and $\overline{\boldsymbol{D_C}}$ unique to each foothold sequence*.* The walker is complementary to all three domains, with nearly identical free energy of hybridization to $\overline{\boldsymbol{D_A}}$ and $\overline{\boldsymbol{D_C}}$; thus, in a field of both $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$, the walker is expected to alternate between binding to each of the footholds, resulting in an indefinite number of steps over the field of footholds.

To mechanistically characterize and optimize the stepping behavior of the transporter

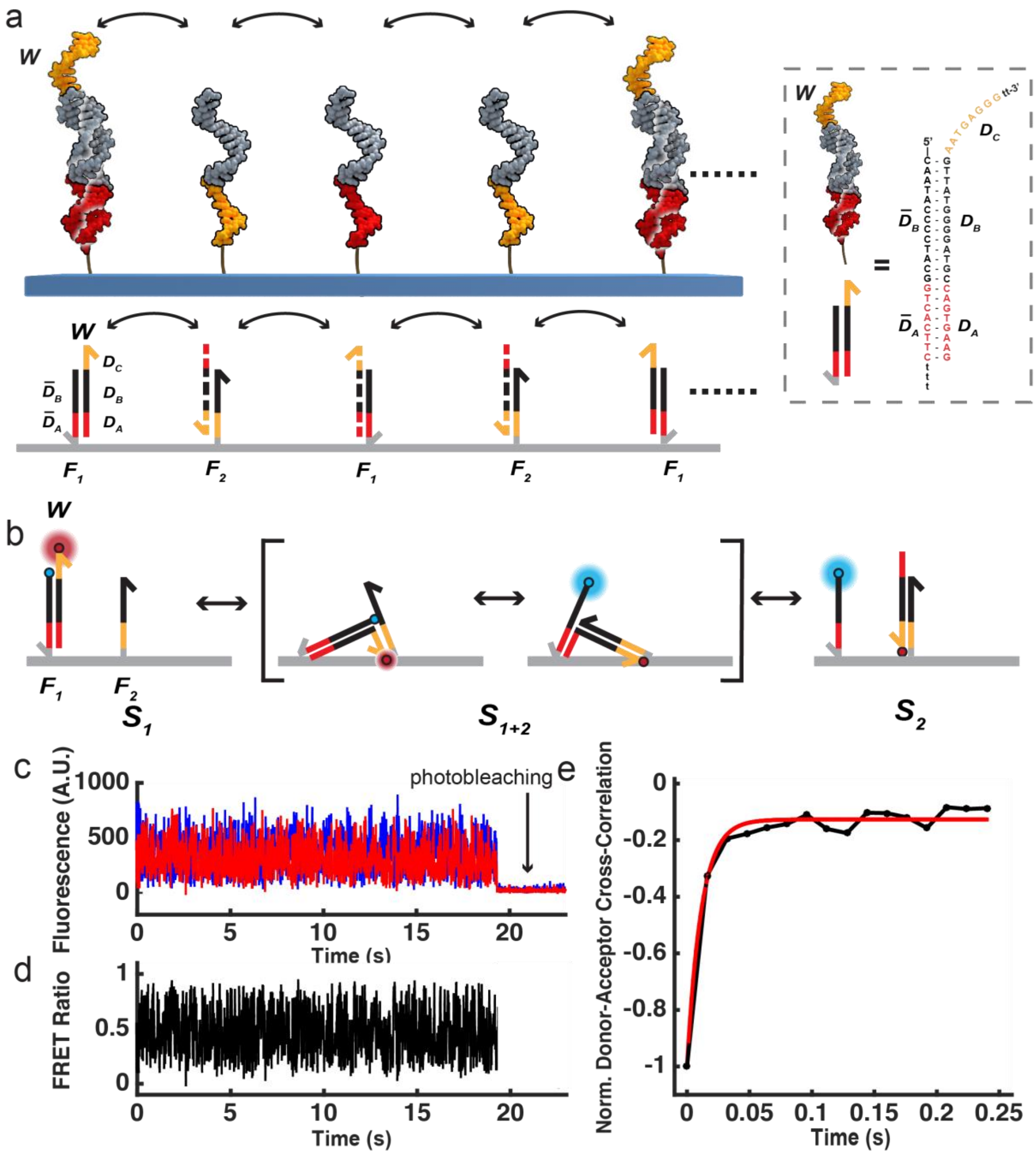


Figure 2.1. Principle and mechanism of a cartwheeling DNA walker. **a**, Schematic showing the intended mechanism of locomotion. **b**, Schematic of smFRET measurement of stepping kinetics of a Cy5-labeled DNA walker on a 2-Foothold DNA tile in which one foothold is labelled with Cy3. **c**, Rapid anti-correlated fluctuations in Cy3 (blue) and Cy5 (red) fluorescence intensity for a single walker-tile complex, suggesting branch migration in hybrid state S1+2. d, FRET ratio versus time for the trajectory shown in c. e, Cross-correlation analysis of Cy3 and Cy5 signal in c, with a single-exponential fit indicating an anti-correlation time constant of 11.4 ms for this trajectory.

, we began with a simple DNA tile system bearing two adjacent footholds for the walker to step between. This ***2-Foothold*** system consists of a 4-helix DNA tile decorated with single-stranded DNA overhangs $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ as well as a biotin label for surface immobilization in TIRF measurements (Figure A1.1). The tile was assembled by thermal annealing of synthetic DNA oligonucleotides and characterized by native polyacrylamide gel electrophoresis (PAGE) (Figure A1.2, A1.3). The footholds are spaced by ~7 nm to match the 17-21 nucleotide length of the duplex formed by binding of the walker to a foothold, and equipped with $(dT)_3$ linkers to provide conformational flexibility. The DNA walker $\boldsymbol{W_{a\_b\_c}}$ consists of a middle branch migration domain $\boldsymbol{D_B}$ of length ***b*** (generally 13 nucleotides) flanked by two toehold domains $\boldsymbol{D_A}$ and $\boldsymbol{D_C}$ (Figure A1.1) with lengths ***a*** and ***c*** (a = c = 5 to 8 nucleotides). The sequence of the walker was chosen to allow for comparison with prior studies of toehold exchange kinetics[28]. To permit measurement of stepping kinetics by single-molecule fluorescence resonance energy transfer (smFRET)[72,73], the 5′-end of $\boldsymbol{F_1}$ is labeled with the FRET donor Cy3, and the 3′-end of $\boldsymbol{W_{a\_b\_c}}$ is labeled with FRET acceptor Cy5. Thus, any stepping of ***W*** between $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ through toehold exchange is expected to give rise to a time-dependent change in FRET efficiency between Cy3 and Cy5 (Figure 2.1b). To reduce the likelihood of analyzing multiple walkers bound to one DNA tile, we combined the walker with a 1.5-fold molar excess of DNA tile, and our smFRET analysis filtered out complexes with >1 photobleaching step in the donor or acceptor channel.

First, we characterized by smFRET the behavior of a walker bearing 8-nucleotide toehold domains ($\boldsymbol{W_{8\_13\_8}}$). While < 30% of walkers exhibited static high- or low-FRET efficiency behavior, which may result from only one of the footholds being present or accessible, the largest fraction of walkers (60~70%) exhibited a FRET efficiency varying between 0.3 and 0.7 (Figure 2.1c, 2.1d). On close inspection, this mid-FRET state contained rapid anti-correlated fluctuations of Cy3 and Cy5 fluorescence intensity, indicating rapid changes in the distance between Cy3 and Cy5 (Figure 2.1e). The cross-correlation function between Cy3 and Cy5 decays with a time constant of 12 ± 3 ms. Due to the rapidity of these fluctuations and the fact that the apparent FRET efficiency occupies a continuum of values rather than discrete states, we hypothesized that these fluctuations are predominantly due to reversible branch migration of domain $\boldsymbol{D_B}$, with only transient dissociation of either toehold from its respective foothold strand. That is, instead of occupying either state $\boldsymbol{S_1}$ or $\boldsymbol{S_2}$ (Figure 2.1b), the walker exists primarily in a hybrid dynamic equilibrium state $\boldsymbol{S_{1+2}}$ in which it is partially base-paired to both $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$. This is consistent with expectations based on the local effective concentration of 250 µM measured in a similar tile-based system[29], where the equilibrium is expected to strongly favor hybridization of the 8-nucleotide toehold sequences. The toehold-dissociated states $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ are not typically observed because of their low occupancy and short lifetimes. Changing the toehold length to 7-, 6-, or 5-nucleotides gives similar kinetic behavior, consistent with the presence of the same branch migration domain length $\boldsymbol{b}$ = 13 (Figure A1.4). Monte Carlo simulations of branch migration within a 13-nucleotide domain (Figure A1.5) suggest that an anti-correlation time constant of ~12 ms will occur when the lifetime of an individual base pair step along the duplex is ~100 µs, which is similar to previous estimates based on kinetic modeling from bulk fluorescence measurements[74] and from three-stranded branch migration of genomic-length DNA sequences[75,76].

To further test our hypothesis of a hybrid ***S_{1+2}*** state, a third foothold strand ***F_1'*** with the same sequence as ***T_1*** was added to assemble a ***3-Foothold*** DNA tile (Figure A1.1, A1.6). The addition of this third foothold is expected to enable the walker to occupy a second hybrid state ***$S_{2+1'}$*** (Figure 2.2a), resulting in a new low-FRET state in addition to the mid-FRET state observed for the ***2-Foothold*** system. Indeed, most mid-FRET trajectories for ***$W_{8\_13\_8}$*** showed slow transitions to and from an additional FRET state of ~0.3 on the 3-Foothold tile (Figure 2.2b), suggesting a new, slower process limited by toehold dissociation. Based on these results, we predicted that decreasing the length of the toehold would yield dramatically faster stepping behavior of walkers, since the rate of dissociation of short DNA duplexes increases exponentially with decreasing length[77,78]. Indeed, walkers with 7-, 6-, and 5-nucleotide toeholds showed two-state FRET behavior with much more rapid transitions between states (Figure 2.2b-i). The median lifetimes (calculated from the per-molecule mean of both high- and low-FRET state lifetimes) decrease from 31.3 s (interquartile range, or IQR, 52.1-16.5 s) for ***$W_{8\_13\_8}$*** to 1.4 s (IQR 3.7-1.1 s) for ***$W_{5\_13\_5}$***, yielding rate constants of stepping that range from 0.03 to 0.72 s^{-1}, or 1.8 to 43 min^{-1} (Figure 2.3a). Given the step size of 7 nm, the fastest walker has an average (undirected) translocation speed of ~300 nm min^{-1}. While the stepping rate increases more than an order of magnitude when the toehold length is decreased to 7 nucleotides, further increases are marginal, and the stepping rate of ***$W_{5\_13\_5}$*** is very similar to that of ***$W_{6\_13\_6}$***. In contrast, kinetic Monte Carlo simulations of stepping on a 3-Foothold tile predict an exponential decrease in the stepping time as a function of toehold length, in direct proportion to toehold dissociation rate constants (Figure A1.5). One possible explanation for this discrepancy is that, as toehold

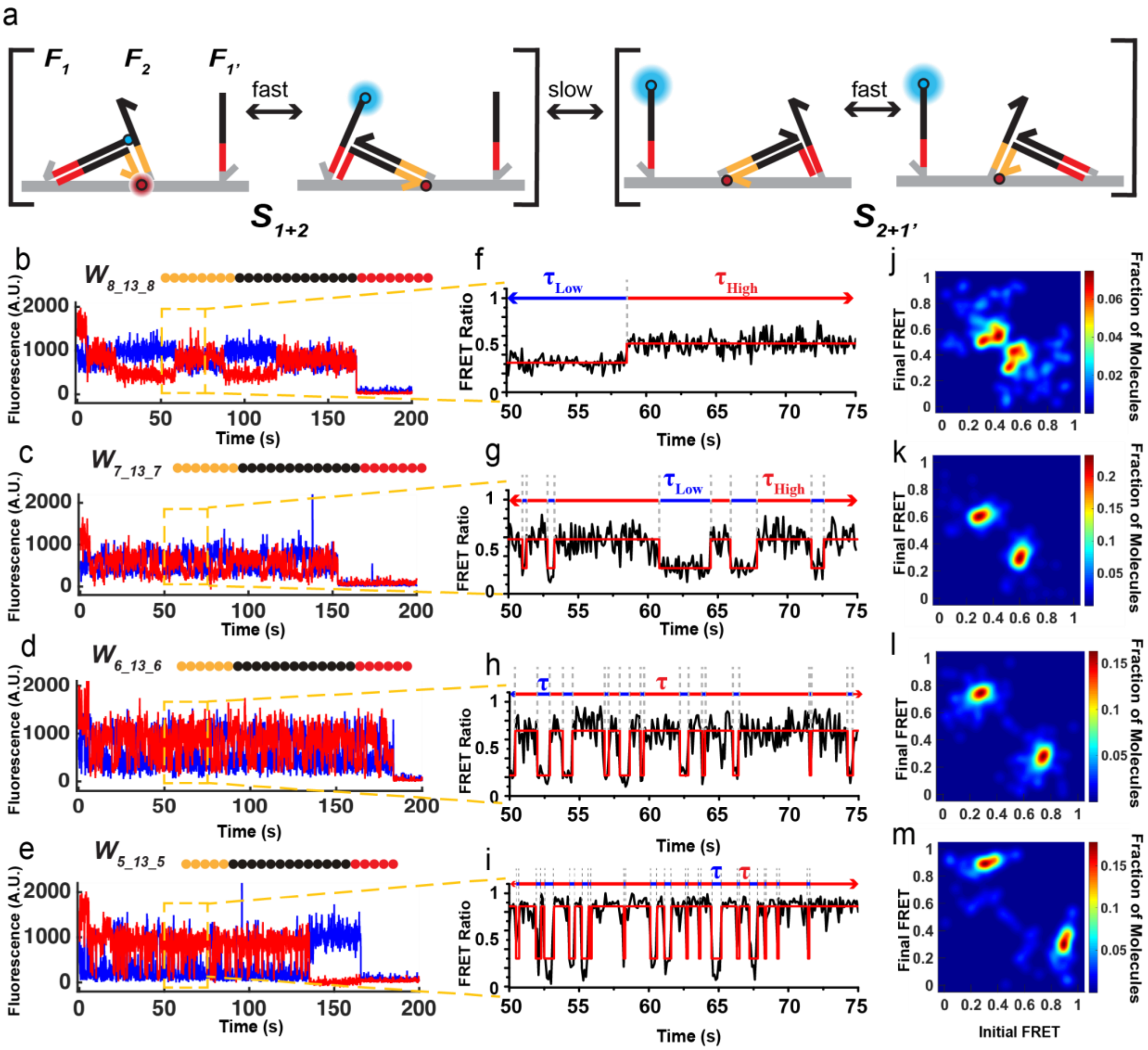


Figure 2.2 Single-molecule FRET characterization of walkers with varying toehold lengths. **a,** Kinetic model of stepping and associated FRET transitions for Cy5-labeled DNA walker on Cy3-labeled ***3-Foothold*** DNA tile. **b**-**e**, Representative single-molecule FRET trajectories of walkers with varying toehold length on a ***3-Foothold*** DNA tile. Cy3 fluorescence is shown in blue, while Cy5 fluorescence is shown in red. The elevated Cy5 signal in the first ~10 seconds of each trace results from direct excitation at 640 nm to confirm the presence of an active acceptor on each walker-tile complex. **f**-**i,** Zoomed-in trajectories showing FRET transitions for 25-s segments of the molecules depicted in **b**-**e**. **j**-**m,** Transition occupancy density plots (TODPs) illustrating the most common FRET transitions for each walker. *N* = 87, 96, 107, and 109 for $W_{5\_13\_5}$, $W_{6\_13\_6}$, $W_{7\_13\_7}$, *and* $W_{8\_13\_8}$, respectively.

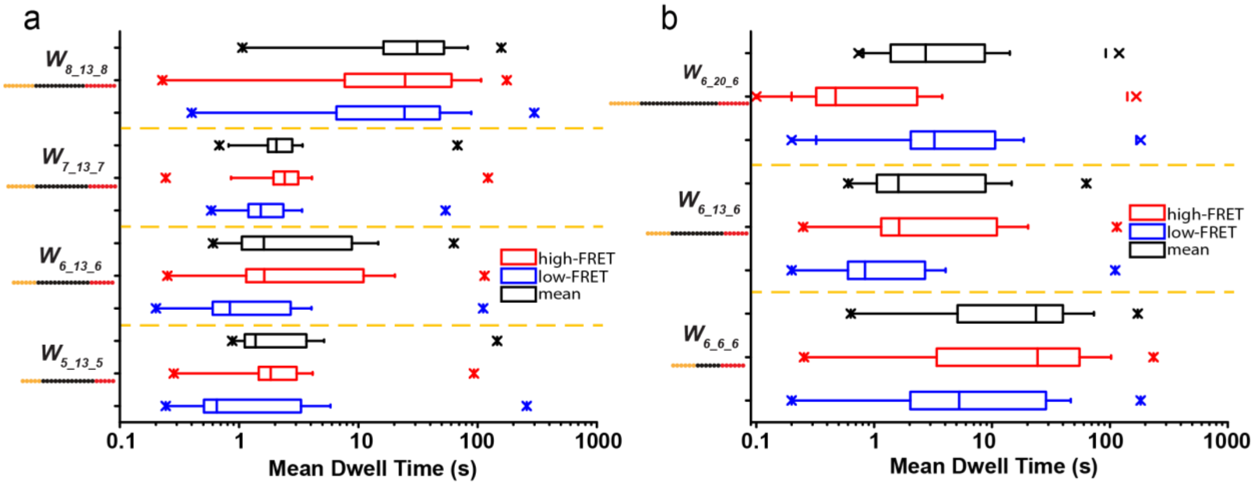


Figure 2.3 Dwell time distribution of walkers with various length of toehold and middle domain. **a,** Box-and-whisker plot of stepping kinetics in the high- and low-FRET states for walkers with varying toehold domain (***D**A* and ***D**C*) lengths. **b,** Box-and-whisker plot for walkers with varying middle domain (***D**B*) lengths. *N* = 105 and 132 for ***W**6_6_6*, ***W**6_20_6*. The box includes the population of all molecules from 25th percentile to 75th percentile; whiskers correspond to 0^{th} and 100^{th} percentiles, excluding outliers. Crosses denote the lower and upper bounds, inclusive of outliers.

nucleotides are removed from the walker, base pairs in the rigid walker-foothold duplex are replaced by unpaired nucleotides near the base of the foothold, which may influence binding and/or dissociation kinetics of toehold $\boldsymbol{D_A}$ by virtue of the much smaller persistence length of ssDNA[51]. For instance, our simulations suggest that a faster association rate of toehold $\boldsymbol{D_A}$ to $\boldsymbol{F_1}$ than $\boldsymbol{F_1'}$ (e.g., due to unintended topological features of the DNA tile that might position $\boldsymbol{F}_2$ at slightly different distances from $\boldsymbol{F_1}$ and $\boldsymbol{F_1'}$), combined with the finite time resolution of our measurements, could yield experimentally measured dwell times that deviate from the predicted exponential dependence in a way that strongly resembles our smFRET observations (Figure A1.5). This is because a toehold $\boldsymbol{D_A}$ might dissociate from, and re-associate with, $\boldsymbol{F_1}$ several times before binding to $\boldsymbol{F_1'}$, giving the appearance of a single long-lived high-FRET state because the unbound state of the toehold is too short for us to resolve experimentally. Consistent with this hypothesis, smFRET measurements do reveal an increasing bias towards high-FRET states as toehold length decreases (Figure 2.2b-e, 2.3a), a bias that is intriguingly reversed when the DNA tile is replaced by a DNA origami substrate. In addition, as the length of toeholds decreases, the difference in apparent FRET efficiency between the two main states increases (Figure 2.2j-m). The observation is consistent with the expected decrease in the distance between the donor and acceptor dyes in the $\boldsymbol{S_{1+2}}$ state, as well as an increase in the donor-acceptor distance in the $\boldsymbol{S_{2+1'}}$ state, when a shorter toehold is present.

To investigate the role of branch migration in the stepping kinetics of walkers with shorter toeholds, we performed smFRET measurements on walkers with different lengths of domain $\boldsymbol{D_B}$, with accompanying changes in the foothold sequences to maintain complementarity with each walker. As with previous walker designs, smFRET transitions were observed for $\boldsymbol{W_{6\_6\_6}}$ and $\boldsymbol{W_{6\_20\_6}}$ (Figure A1.7). Kinetic Monte Carlo simulations predict that stepping dwell times will increase linearly in proportion to the length of $\boldsymbol{D_B}$, since walkers will spend a larger fraction of their time in branch migration intermediates from which toehold dissociation is difficult or impossible. Indeed, when the middle domain increases to 20 nucleotides (walker $\boldsymbol{W_{6\_20\_6}}$), the median stepping lifetime is 2.7 s (IQR 8.7-1.4 s), which is slightly longer than that of $\boldsymbol{W_{6\_13\_6}}$ (1.6 s, IQR 8.7-1.1 s). However, when $\boldsymbol{D_B}$ is decreased to 6 nucleotides (walker $\boldsymbol{W_{6\_6\_6}}$), the median stepping lifetime increases to 23.6 s (IQR 39.4-5.2) (Figure 2.3b), contrary to the predictions of a simple branch migration model, and again suggesting that the structural details of the walker-tile complex (such as the match between walker length and foothold spacing) may play a role. While the $(dT)_3$ ssDNA spacers between the foothold strands and the tile are expected to provide sufficient flexibility to compensate for the difference of ±2.3 nm from the addition or subtraction of 7 nucleotides from $\boldsymbol{D_B}$, altering the length of the branch migration domain may still introduce an incongruity between the reach of the walker and the spacing between foothold strands, since the distance between adjacent foothold strands is fixed at ~7 nm. In any case, the lack of a positive correlation between the length of $\boldsymbol{D_B}$ and the stepping lifetime suggests that any impact of branch migration upon the stepping rate for values of ***b*** between 6 and 20 nucleotides is overshadowed by other factors, such as toehold binding and dissociation kinetics and the match between walker length and foothold spacing.

2.3.2 Long-range walker movement on 2D foothold arrays

Based on smFRET measurements, the DNA walker with the fastest stepping rate and most homogeneous behavior is ***$W_{6\_13\_6}$.*** To test the performance of this optimized DNA walker as a long-distance 2D transporter, we developed a surface modification method that yields a high density of DNA footholds on a glass coverslip. In this method, alkyne-functionalized ***F_1*** and ***F_2*** are attached to an azide-modified coverslip through copper-free click chemistry, resulting in random, high-density conjugation of the two different foothold strands to the surface (Figure 2.4a, b; also see section 2.2 Materials and Methods). Using TIRF microscopy, we measured an average oligonucleotide surface density of 1.67×10^4 /μm^2, which is predicted to yield distances between nearest-neighbor ***F_1*** and ***F_2*** strands varying from ~2-13 nm, assuming a completely random distribution of footholds on the slide surface (Figure 2.3c, Figure A1.8, and section 2.2 Materials and Methods). This estimate should be interpreted as an upper bound on the average spacing between footholds, since our measurements may underestimate the true density if a significant fraction of probe fluorophores are photobleached prior to the measurement. To confirm that larger inter-foothold distances are compatible with rapid stepping by the walker, we repeated our smFRET measurements of stepping in a ***3-Foothold*** system constructed from a distinct DNA origami scaffold (Figure A1.9, Table A1), using a dT_6 linker instead of a dT_3 linker between the footholds and the origami for added flexibility. Despite the larger distance between adjacent foothold sites on this DNA origami (10.44 nm on average, assuming 0.33 nm/nucleotide) compared to the previous tile system (~7 nm), smFRET characterization revealed a similar stepping rate constant of ~0.5 s^{-1} (IQR 0.6-0.1 s^{-1}) for ***$W_{6\_13\_6}$*** on this

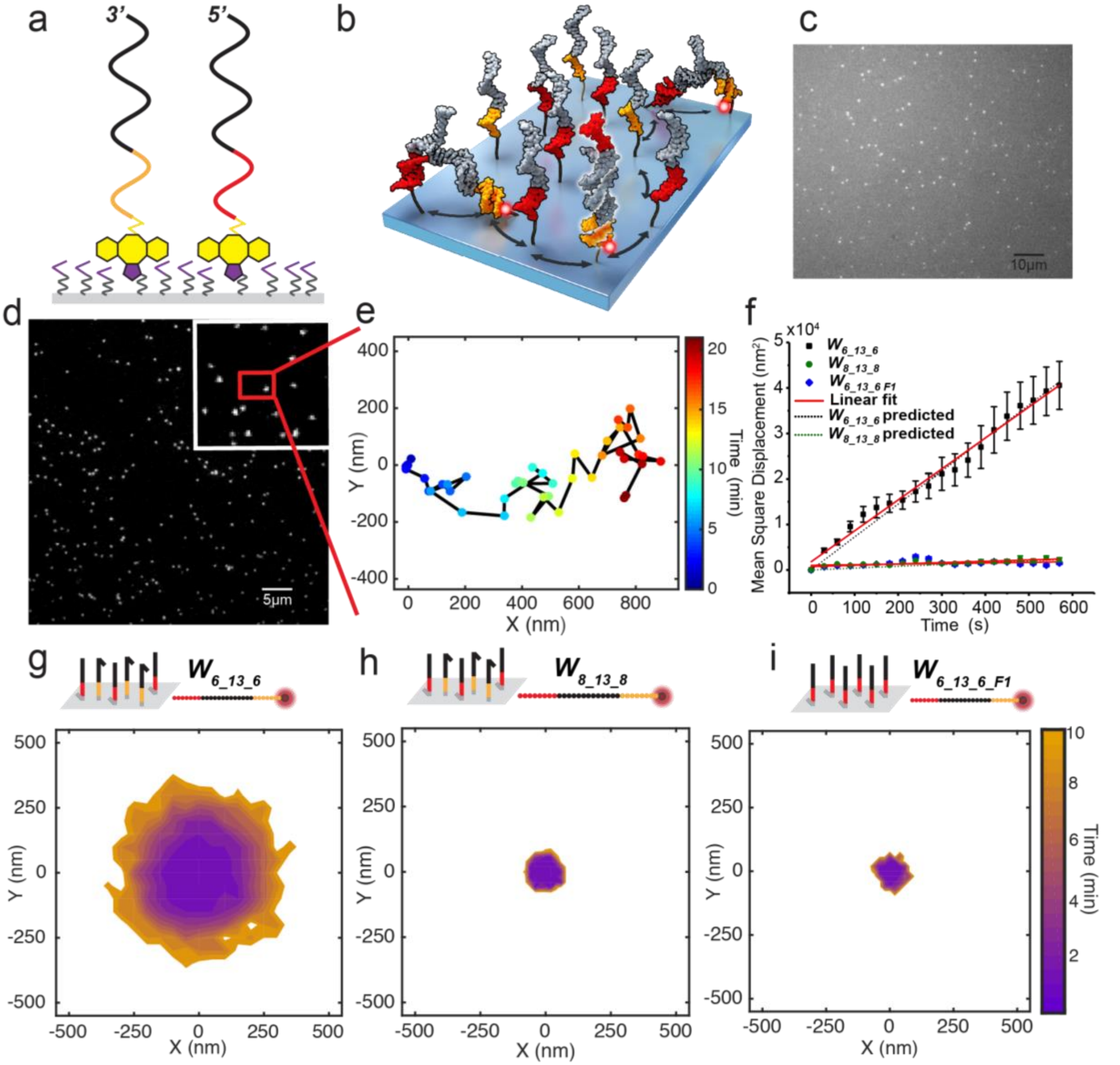


Figure 2.4. Characterization of 2D foothold arrays and long-range walker movement. **a,** Schematic of ***F_1*** and ***F_2*** DNA conjugated to glass coverslip surface at high density *via* copper-free click chemistry. Surface azides are shown in purple and DBCO in yellow. **b,** Schematic of walker movement over a 2D array of footholds. **c,** TIRF image of complementary oligonucleotides bound to footholds on coverslip surface at a ratio of 1 fluorescently labelled oligo:1 million unlabelled fluorescent oligonucleotides. **d-e,** TIRF image of Cy5-labeled *$W_{6\_13\_6}$* on ***F_1*** and ***F_2*** coated quartz slide (**d**) and one representative fast-moving trajectory of *$W_{6\_13\_6}$* (**e**). **f**, Mean square displacement (MSD) *versus* time plot for ***$W_{6\_13\_6}$**,* ***$W_{8\_13\_8}$**,* and ***$W_{6\_13\_6}$*** control (***F_1*** only). Error bars represent one standard deviation. The diffusion coefficients derived from linear regression fits to the MSD *versus* time data are 17, 2.3, and 0.33 nm^2/s, respectively. The MSD curves predicted from the stepping kinetics measured by smFRET on DNA tiles (dotted lines in green and black) are also shown for comparison. **g-i**, Comparison of extent of diffusion (region comprising 95% of trajectories, with all starting at the origin) of ***$W_{6\_13\_6}$**,* ***$W_{8\_13\_8}$*** and ***$W_{6\_13\_6}$ F_1***-only control over a 10-minute period of observation.

new scaffold, suggesting that the stepping rate is robust to small perturbations in foothold spacing.

Next, we characterized the long-range movement of optimized walker ***W**$_{6\_13\_6}$* by 2-dimensional single-particle tracking using TIRF microscopy (Figure 2.4d-e). Most of the molecules travel >200 nm from their starting position within 10 minutes (Figure 2.4f), resulting in a measured 2D diffusion coefficient of 17±0.5 nm^2 s^{-1} (R^2= 0.99), and some molecules are observed to travel nearly 1 µm before photobleaching (Figure 2.4d). In contrast, particle tracking of ***W**$_{8\_13\_8}$* indicates a much smaller diffusion coefficient of 0.7 ± 0.1 nm^2 s^{-1} (R^2= 0.73) (Figure 2.4f, h), or 2.2±0.2 nm^2 s^{-1} (R^2= 0.89) if a single fast-moving outlier is included (Figure A1.9), consistent with predictions based on its ~10-fold slower stepping rate as measured by smFRET. For comparison, a random walk model with step sizes of 10.8 and 11.1 nm for ***W**$_{6\_13\_6}$* and ***W**$_{8\_13\_8}$*, respectively (Figure A1.8), and stepping rates taken from smFRET measurements on 3-Foothold DNA tiles predicts diffusion coefficients of 18.1 nm^2 s^{-1} and 0.99 nm^2 s^{-1} for ***W**$_{6\_13\_6}$* and ***W**$_{8\_13\_8}$*, respectively. This close agreement with predictions from measurements of stepping kinetics and foothold density suggests that the optimized walker functions as designed, even on a different substrate and over substantially longer distances. Moreover, a surface coated with only a single foothold type (***F**$_1$*) results in no significant diffusion of ***W**$_{6\_13\_6}$* (D ~ 0.33±0.20 nm^2/s (R^2= 0.09), standard deviations in position σ_x = 16.4 nm, σ_y = 15.2 nm), indicating that the observed diffusion is indeed the result of the designed walking mechanism involving both footholds (Figure 2.4f, i; Figure A1.12).

2.4 Discussion

We have created a new class of cartwheeling single-stranded DNA walker that exploits a toehold exchange mechanism to traverse arrays of specific oligonucleotide sequences in a cartwheeling fashion. The present walker's directionally unbiased movement has useful precedents in both nature and nanotechnology. For example, the kinesin MCAK utilizes undirected, one-dimensional diffusion to rapidly locate the ends of microtubules for depolymerization, resulting in faster searching over short distances than would be possible with direct binding from solution[16]. In the field of nanotechnology, synthetic biochemical cascades have exploited undirected, two-way transport to promote reagent channeling between coupled enzymes using a swinging arm over short distances (~10 nm)[29], and a cargo-sorting robot was recently reported to use unbiased diffusion to transfer payloads over distances of tens of nanometers in a period of hours[57]. It is likely that the DNA acrobat's "cartwheeling" mode of locomotion plays an important role in generating rapid stepping relative to similar systems studied previously, since some of these employed similar domain lengths and yet still exhibited stepping orders of magnitude slower than our system[57]. One advantage of the cartwheeling geometry is that rigid double-stranded segment always bridges between adjacent footholds, ensuring rapid toehold binding as long as there is a good match between walker length and foothold spacing. Secondly, Thubagere *et al.*[57] suggest that their cargo-sorting robot may exhibit slow branch migration when strand displacement is initiated at the distal end of a foothold and proceeds toward the surface, due to the entropic cost of stretching the single-stranded DNA away from the surface; if this is true, the DNA acrobat overcomes this issue by virtue of the fact that branch migration always proceeds away from the point of attachment.

The present study of a cartwheeling DNA walker also shows that the speed and range of similar DNA-based systems may be improved with careful optimization. The fastest of our toehold exchange walkers can search among ~43 foothold sites per minute with a stepping distance of ~10 nm. While still much slower than many natural motor proteins (*e.g.,* 0.38 μm^2 s^{-1} for MCAK[16]), the stepping rate of this cartwheeling walker is more than an order of magnitude higher than that of other DNA-only walker systems. This improvement in performance was enabled by detailed single-molecule analysis of stepping kinetics as a function of key design parameters, an approach that is likely to be generalizable to many other systems in nanotechnology. While decreasing the toehold length from 8 to 5 nucleotides yields faster stepping rates as predicted, the marginal improvements in stepping rate appear to diminish below ~6 nucleotides, in contrast to the predictions of our kinetic modeling; indeed, a walker with a 4-nucleotide toehold exhibited no evidence of stepping behavior at all by smFRET (data not shown). These results, as well as those for the shortened walker $W_{6\_6\_6}$, suggest that optimizing the mechanical properties of the system (reach of the walker, entropic tension in single-stranded linker segments, etc.) may also be important, and could yield further improvements in the future.

Finally, the present characterization of toehold exchange reactions at very high local effective reagent concentrations in a variety of contexts suggests that it may be challenging to obtain rate constants significantly faster than 1 s^{-1} for conventional strand displacement operations in DNA nanomachines. To surpass this apparent “speed limit”, dynamic DNA nanotechnology may need to incorporate further innovations inspired by natural systems, such as more precise control of local DNA mechanics and conformational changes, as well as judicious coupling to (rapid) exergonic processes.

Chapter 3:

Nucleic Acid Detection by Single-Molecule Kinetic Fingerprinting[3, 4]

3.1 Introduction

Due to limitation of using the polymerase chain reaction (PCR) amplification in most nucleic acid analysis, which has been introduced in Chapter 1.3, several amplification-free methods[80–82] have been pursued for the analysis of nucleic acids and other biomolecules, in some cases permitting the direct capture and quantitation of analytes from biological matrices without prior purification. However, these amplification-free approaches typically suffer from a different set of challenges. First, since they lack the geometric amplification of PCR, these methods are generally limited by finite thermodynamic discrimination factors between closely related sequences [83]. This thermodynamic specificity limit is embodied by the parameter $Q_{max,therm} = e^{\frac{-\Delta\Delta G^{o}}{RT}}$, where $\Delta\Delta G^o$ is the difference in the Gibbs free energy of hybridization of a detection probe to a target sequence and of the same probe to a related but spurious target sequence; in practice, this translates to $Q_{max,therm}$ values ranging from about 20 to

[3] Reproduced in part from Johnson-Buck, A., Li, J., Tewari, M., & Walter, N. G. Methods, 153, 3-12. Copyright Elsevier, 2018. First two authors contributed equally to this work.

[4] Alexander Johnson-Buck performed smFRET measurement of miR-16 and L858R. Jieming Li reviewed the SiMREPS methods. Jieming Li performed experiments on DNA methylation. Jieming Li and Alexander Johnson-Buck co-wrote the paper.

20,000 for single-nucleotide variants [83]. In most cases, the actual single-base specificity realized is only 90-99%[84,85]. Second, since many amplification-free assays are surface-based, true single-molecule sensitivity becomes challenging due to the inability to completely suppress nonspecific binding of probes to the detection surfaces.

To realize amplification-free biomolecule detection without being bound by thermodynamic limits of specificity, we developed an approach based on time-resolved measurement of the interaction kinetics between fluorescent probes and single immobilized analyte molecules [86]. This approach, termed SiMREPS (single-molecule recognition through equilibrium Poisson sampling), exploits repeated observations of transient probe interactions with each surface-bound copy of the analyte to create a "kinetic fingerprint" that is highly characteristic of that particular analyte molecule when detected at the single molecule level (Figure 3.1a), and is significantly perturbed by even small alterations such as single-base substitutions. As a result, nonspecific binding of probes to the surface and to closely related sequences can be confidently screened out due to their distinct kinetics (Figure 3.1b, c), yielding essentially background-free detection of single analyte molecules after applying appropriate filters for signal-to-noise, intensity, and probe binding and dissociation kinetics (Figure 3.1d). To facilitate the observation of repeated fluorescent probes binding to the same copy of analyte, the analyte is typically immobilized to a biotin-functionalized surface *via* a streptavidin bridge and a biotin-labeled capture probe (Figure 3.1a). While DNA oligonucleotides have been successfully employed as capture probes for SiMREPS, several locked nucleic acid (LNA) modifications are usually incorporated when the analyte is a short nucleic acid such as a miRNA to permit high-affinity capture while leaving several unpaired nucleobases to interact with the fluorescent probe.

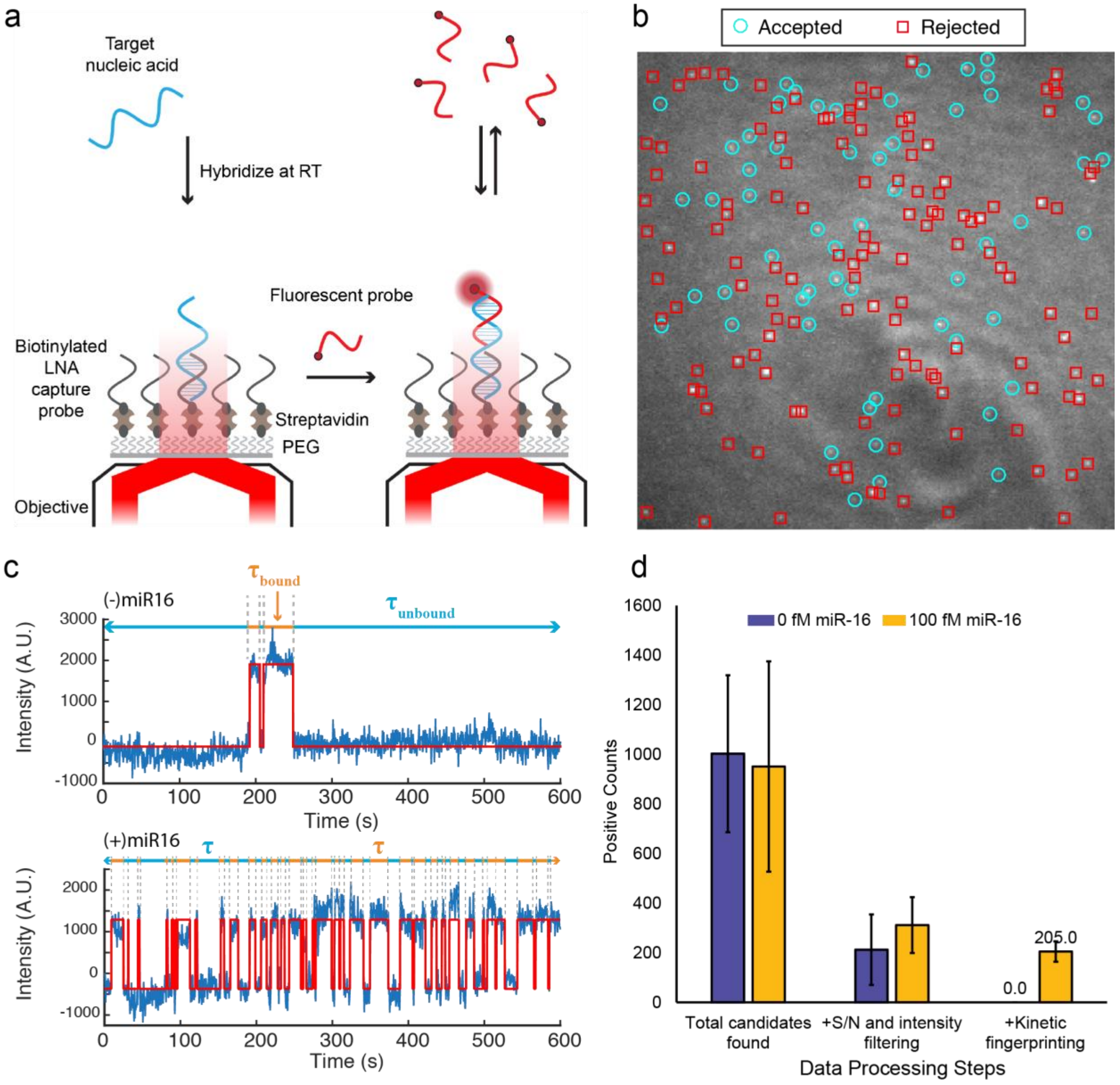


Figure 3.1. Overview of the SiMREPS technique for low-background, high-specificity detection of single molecules. **a**, Schematic illustrating the experimental principles of SiMREPS. A target analyte is captured at the surface of a coverslip *via* a biotinylated capture probe. Then, using TIRF microscopy, each copy of surface-bound analyte is detecting by monitoring the repeated transient binding of a fluorescent probe, which yields a distinctive kinetic fingerprint; **b**, Single movie frame from a representative field of view from SiMREPS using objective-type TIRF microscopy. Red squares indicate positions of binding events that were rejected as likely background binding by kinetic fingerprinting, and the cyan circles indicate positions of repeated binding events with kinetics that suggest the presence of the analyte; **c**, Representative fluorescence-*versus*-time traces observed in the presence and absence of a miRNA target, *hsa*-miR-16. The kinetics of transitions between FP-bound and FP-unbound states are analyzed to distinguish between true and false positives at the single-molecule level. **d,** Number of spots counted in positive and negative control experiments for miR-16 before ('total counts') and after ('accepted counts') kinetic filtering. While

filtering based on intensity and signal-to-noise (S/N) alone does not yield a significant difference between positive and negative controls (due to background binding of the probe), the application of kinetic filtering criteria (see section 2.7.4) reduces accepted counts in the negative control to essentially zero.

Because the binding of fluorescent probes to a single analyte molecule can be modeled as a Poisson process, the number of probe binding and dissociation events observed for each analyte molecule (N_{b+d}) will increase linearly over time, with a coefficient of variation (C.V.) that decreases as $\sim\frac{1}{\sqrt{N_{b+d}}}$ [86]. This decrease in C.V. with increasing observation time permits the kinetic fingerprint resulting from a single analyte molecule to be separated from the signals resulting from nonspecific binding to an *arbitrarily high* degree. Similarly, the lifetimes of the analyte in the probe-bound (τ_{bound}) and probe-unbound (τ_{unbound}) states become better separated from the background binding as an increasing number of probe-binding events to each analyte is observed. This increased confidence in the source of a given kinetic fingerprint is the core feature of SiMREPS, and means that probes with finite thermodynamic discrimination can be used to detect an analyte with arbitrarily high specificity, given an adequate number of binding events. In other words, the specific time evolution of the detection signal becomes a heretofore untapped observation parameter that serves to enhance the accuracy of analyte identification, in concept similar to the revolution conventional fluorescence microscopy experienced upon introduction of super-resolution approaches that observe a time series of sparse signals from single molecules to determine their cellular localization more accurately [87]. As a proof of concept, we show that miRNAs such as miR-16 [86] can be detected using SiMREPS with essentially zero background signal from surface binding of fluorescent probes if kinetic fingerprints from single molecules are filtered by N_{b+d} and τ_{bound} (Figure 3.1d).

In this chapter, I will discuss practical considerations for the use of SiMREPS to detect short nucleic acid such as miRNA and DNA fragments, including guidelines for instrumentation and assay design. Upon that, we demonstrate the high specificity of the technique through proof-of-concept measurement of the cancer point mutation *EGFR* L858R with an apparent discrimination factor of > 1,000,000. In addition, I will discuss the possibility of using SiMREPS to detect DNA CpG methylation. Besides analytes scope expanding, I will also introduce some studies on method optimization of SiMREPS.

3.2 Materials and Methods

3.2.1 Instrumentation and sample cell design

Since SiMREPS in its current implementation requires the presence of an excess of fluorescent probe in binding equilibrium with the surface-immobilized analyte, a microscope capable of total internal reflection fluorescence (TIRF) illumination is required to reject background signal from the majority of freely diffusing (non-surface-bound) fluorescent probes. Most commonly, TIRF measurements are carried out using either a prism-type (P-TIRF) or objective-type (O-TIRF) illumination geometry (Figure 3.2a, b). Excitation light is provided by a laser of appropriate wavelength (e.g., 640 nm for probes labeled with Cy5) and output power (typically 10-100 mW) and undergoes total internal reflection at the interface between the coverslip and the aqueous solution containing the fluorescent probe. To reliably detect single fluorescent probes with satisfactory signal-to-noise, an illumination intensity of ~50 W/cm^2 is typically used, and the TIRF angle adjusted to achieve a calculated penetration depth of ~80-110 nm of the evanescent field. Emission light from surface- or analyte-bound fluorescent probes is collected through a microscope objective lens, passed through dichroic mirrors and/or

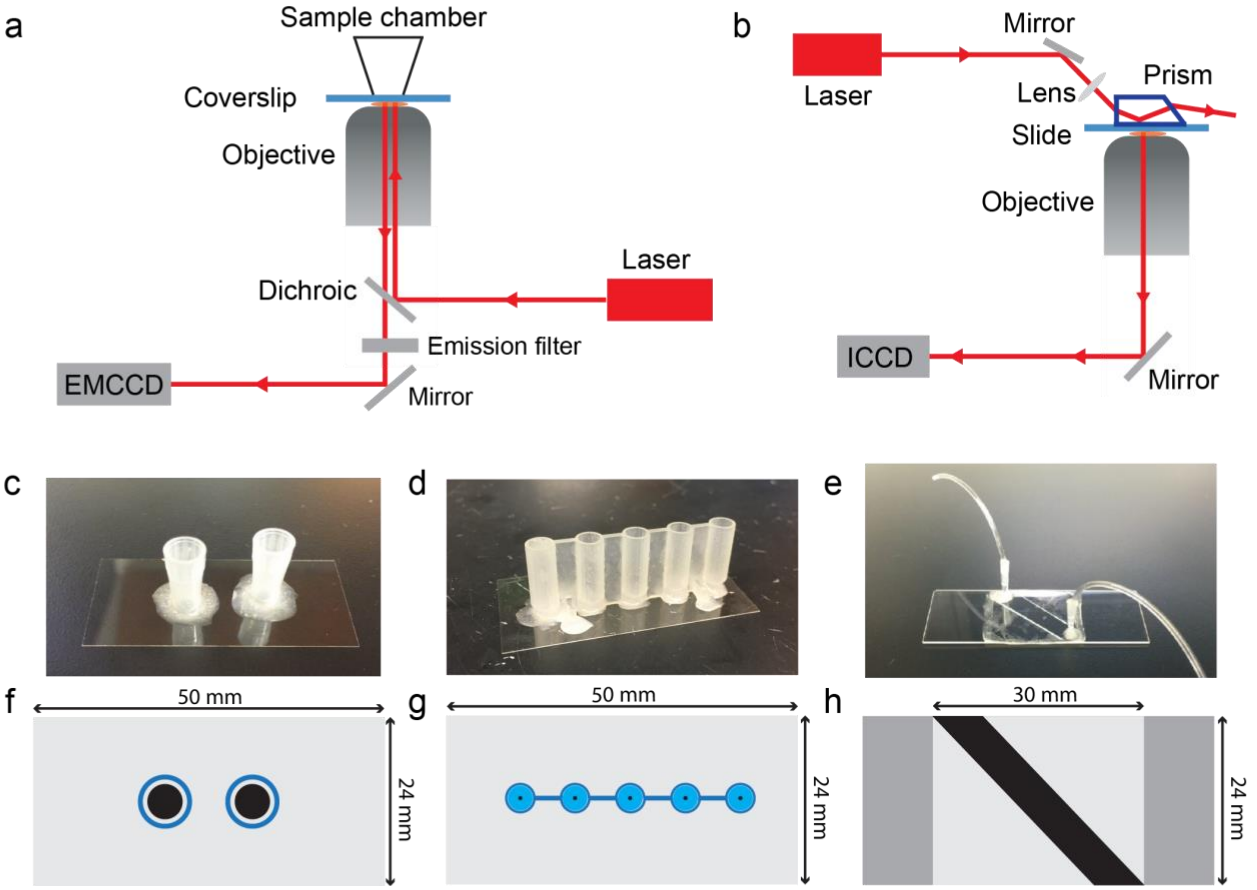


Figure 3.2. Overview of instrumentation and sample cells. **a**, Objective-type TIRF microscope. **b**, Prism-type TIRF microscope. **c**, Pipet tip chamber sample cell. **d**, 3D-printed sample cell with cylindrical reservoir and tapered conical base. **e**, Sandwich-type sample cell for prism-TIRF measurements. **f**-**g** Scale drawings showing a top view of each sample cell type shown in **c**-**e**. The black-shaded region in each panel represents the surface area available for target capture and imaging on the coverslip or slide. Blue-shaded regions in **f** and **g** represent the plastic walls of the sample wells.

chromatic filters to remove the majority of the excitation light, and detected by a high-sensitivity camera such as an ICCD, EMCCD, or sCMOS. In our study, an EMCCD camera is used in O-TIRF and an ICCD camera is used in P-TIRF. In SiMREPS imaging, the signal integration time (exposure time) per frame is typically 500 ms, and typically 1200 movie frames are acquired per field of view (FOV).

SiMREPS is compatible with a variety of sample cell types (Figure 3.2c-h). Because the sample cell must be positioned between the prism and objective, P-TIRF requires thin flow cells that are typically constructed by sandwiching two pieces of double-sided tape between a coverslip and a biotin-functionalized microscope slide, with optional plastic tubing added for ease of sample injection (Figure 3.2e). However, with O-TIRF taller sample cells constructed from cut pipet tips (Figure 3.2c) or 3D-printed plastic parts (Figure 3.2d) attached to a biotinylated coverslip may also be used. These taller sample cells permit the immobilization of analyte on the imaging surface at higher densities, providing greater sensitivity than thin flow cells. Thus, for high-sensitivity measurements (LOD < 1 pM) O-TIRF is preferred over P-TIRF for SiMREPS. However, due to their open-top geometry, measurements that take a long time (>1 h) or using fluorescent probes with slow-off rates (< 2 min^{-1}) may benefit from filling the sample chamber to the top with imaging solution and sealing it with parafilm to slow the influx of atmospheric oxygen. All the data presented here were collected by O-TIRF using sample cells constructed from cut pipet tips. Recently, other instrumentation has been introduced for super-resolution studies, including spinning disk confocal microscopes (CSU-W1, Yokogawa Electric) [88] and the Oxford Nanoimager [89]; these may provide other options for SiMREPS measurements in the future.

3.2.2 Analyte Scope

Since it does not require any nucleic acid-specific enzymes such as ligases or polymerases, SiMREPS is in principle capable of detecting any analyte that can (1) be immobilized at a surface, preferably *via* a specific interaction, and (2) remain free to transiently recruit fluorescent probes from solution while bound to the surface. It thus has a much broader scope than amplification-based approaches. To date, SiMREPS has been successfully applied to the identification and counting of short nucleic acids such as miRNAs (miR-16, miR-21, let-7a, let-7c, miR-141, *cel*-miR-139) [86] and ~22-160 bp fragments of single-stranded or double-stranded DNA such as cancer-related *EGFR* mutations (see Results). Since the assay is typically performed at ambient room temperature, to ensure maximal sensitivity for double-stranded or highly structured analytes, care must be taken to fully denature and sequester any interfering secondary structure that might interfere with surface capture or fluorescent probe binding, *e.g.*, by brief heating in an excess (e.g., 1-2 µM) of a carrier oligonucleotide or sequence-specific oligonucleotides that prevent the formation of interfering secondary structure. In contrast, short nucleic acids that are difficult to detect with amplification-based approaches are readily detected by SiMREPS and are thus particularly strong candidates for the technique. Finally, owing to its high specificity, SiMREPS is capable of discriminating single-nucleotide variants such as let-7a and let-7c [86].

3.2.3 Probe design

3.2.3.1 Capture probes

For sequence-specific capture of analytes, terminally biotin-labeled capture probes (CPs) are immobilized on a streptavidin-coated coverslip or microscope slide surface.

The CPs may, in principle, comprise any type of nucleotide or modified nucleotide, including DNA, RNA, and other non-natural nucleic acids such as LNAs, peptide nucleic acids (PNAs), or unlocked nucleic acids (UNAs) [90]. When using SiMREPS for miRNA detection, it is important to leave ~10 nucleotides unpaired for interaction with the fluorescent probe, necessitating the use of a relatively short capture probe (10-12 nucleotides). We therefore typically employ CPs comprising mostly DNA nucleotides, but also incorporating several (4-5) LNA nucleotides to maintain a high melting temperature (T_m) despite the short length. Compared to natural DNA and RNA oligonucleotides, an LNA oligonucleotide offers substantially increased affinity for its complementary strand, which makes it an ideal capture probe of short RNA and DNA targets. The positions of LNA modifications are determined semi-empirically using the online T_m and self-structure prediction tools available from Exiqon [91]; the goals are to achieve a predicted T_m (under standard conditions) > 60 °C and a self-structure score as low as possible, preferably < 25 °C. LNA capture probes used in our study were purchased from Exiqon (now distributed by Qiagen) with HPLC purification. For instance, the melting temperature of the capture probe for miR16 increases from 47 °C to 79 °C when replacing 4 out of 10 DNA nucleotides with the corresponding LNA nucleotides. For longer targets such as genomic DNA fragments, long DNA capture probes with suitably large T_m values (> 60 °C) have also been employed with success; these have the advantage of higher capture specificity than short CPs. Regardless of the type of CP used, the choice of capture region should be chosen such that it minimizes any interfering secondary structure in the CP and target. Such optimization can be carried out using prediction tools such as Exiqon's OligoAnalyzer, Integrated DNA Technology's T_m prediction tool, or NUPACK[92–95].

3.2.3.2 Fluorescent probes

To permit kinetic fingerprinting of single molecules by SiMREPS, reversible binding is required to allow for many cycles of binding and dissociation of the fluorescence probe (FP) to each copy of the target. Typically, this means that the T_m of the interaction between the FP and target should be comparable to the temperature at which the assay is conducted (usually room temperature, 20-25 °C). The lifetime of bound state should be longer than the camera exposure time (in our case, 500 ms) but not so long as to impede the observation of enough binding events to separate the positive signal from background binding within a convenient sample imaging time frame (in our case, typically ~10 minutes). At constant temperature and ionic strength, the dissociation kinetics of a short oligonucleotide probe are exponentially dependent upon the length of the probe [96], making the choice of FP length a particularly important parameter [86]. In the high-ionic strength buffers typically used in SiMREPS measurements of nucleic acids (see section3.2.6) and for observations near room temperature, the optimal length of FPs with ~50% GC content is typically ~9 nucleotides for RNA targets, and ~8 nucleotides for DNA targets. Probes against sequences with high GC content can be designed with one or more intentional mismatches to achieve appropriate kinetics; alternatively, denaturants such as 5-30% formamide can be added to mildly destabilize the FP-target interaction. Formamide lowers the (T_m) of DNAs linearly by 2.4–2.9°C/mole of formamide depending on the (GC) composition and state of hydration [97]. Higher observation temperatures (e.g., using a heated microscope objective and/or stage) can be contemplated as another way of destabilizing FP-target interactions. If denaturants or higher temperatures are used, the stability of the CP-target interaction should be verified under the new conditions, e.g., by performing SiMREPS measurements after variable incubation times and determining whether there is a systematic decrease in detected target molecules over time.

When choosing the binding register of the FP on the target sequence, the following criteria should be observed for optimal performance:

1. GC content of the FP-target interaction should be ≤ 50%, if possible, to ensure rapid binding and dissociation kinetics;
2. There should be at least 1-2 unpaired nucleotides between the binding sites of the CP and FP on the target in order to avoid stacking interactions between adjacently binding probes that will tend to lengthen the bound-state lifetime of the FP;
3. It is preferable to position the fluorophore distally on the FP relative to the CP, to reduce the likelihood of stacking interactions between the fluorophore and the CP; alternatively, an additional 1-2 unpaired bases between the FP and CP can accommodate a proximally positioned fluorophore;
4. If single-base discrimination is desired, note that the selectivity is higher when the mismatched nucleotide is near the middle of the FP than it is when positioned near the 3′- or 5′-end of the probe-target duplex. While mismatches near the end of the duplex can also provide adequate discrimination by SiMREPS [86], longer observation times may be necessary to achieve perfect kinetic discrimination.

Notably, the use of fluorescent probes with only 8-9 nucleotides will not provide sufficient specificity to uniquely identify a sequence against a background of genomic DNA or RNA. Additional specificity is provided by the capture probe (~10 nucleotides), which can be engineered to be as specific as needed, for example, by lengthening it upon removal of LNA moieties or increase of the assay temperature or formamide concentration, and the slide surface should be well passivated against nonspecific binding of nucleic acids. Furthermore, addition of a second fluorescent probe to create a FRET pair has been employed in super-resolution imaging with DNA-PAINT [98] and could provide additional specificity by requiring the proximity of two short (e.g., 8-10 nucleotide) sequences to observe a positive kinetic fingerprint. The addition of a second fluorescent probe will slightly increase the footprint of the assay (from ~20 to ~30 nucleotides), but this footprint will still be comparable to, or shorter than, that required by the majority of other nucleic acid assays based on PCR or thermodynamic binding, while also providing extremely high single-base discrimination power without any purification or enzymatic processing. Since SiMREPS has notably fewer required components than enzymatic assays, the choice of both probes and buffer conditions is particularly flexible and can be adjusted to match most specificity requirements imposed by a particular sample matrix.

3.2.3.3 Auxiliary oligonucleotides

While SiMREPS can often be performed using only the CP and FP, other oligonucleotides may be helpful in preventing re-hybridization of double-stranded targets, in preventing secondary structures in the target that could interfere with FP binding, or in reducing off-target binding of the FP to the CP or spurious target sequences.

1. Carrier oligonucleotide: 1-5 µM of a polythymidine oligonucleotide such as $(dT)_{10}$ can reduce sample loss due to adsorption as well as prevent re-hybridization of double-stranded DNA targets after denaturation.
2. CP blocker: some combinations of CP and FP sequences will result in a large amount of transient FP binding to the CP, which can lead to false positives or false negatives; in such cases, a short oligonucleotide probe complementary to the CP can be added to the imaging solution at a sufficient concentration (e.g., > 10 nM) to saturate any non-target-bound CPs at the imaging surface.
3. Competitor oligonucleotides: to block transient binding to closely related sequences, short unlabeled oligonucleotides may be included in the imaging solution. For instance, in the detection of *EGFR* L858R presented in this work, an 8-nucleotide probe complementary to the wild-type (WT) sequence – a so-called WT competitor – is used to reduce binding of the FP to the WT *EGFR* sequence.
4. Secondary structure blockers: short (10-14 nucleotide) oligonucleotide probes complementary to the regions of the target that are directly adjacent to the CP and/or FP binding region can be useful in improving both capture efficiency and accessibility of the target to the FP. These may be added either prior to surface capture or in the imaging buffer.

3.2.4 Slide and sample cell preparation

3.2.4.1 Surface functionalization

The objectives of surface functionalization are twofold: first, to passivate the imaging surface against excessive nonspecific binding of the FP and other components; and second, to provide an affinity tag, usually biotin, that can be used for subsequent immobilization of the CP. Whether glass coverslips or microscope slides are used as the imaging surface, a typical surface functionalization is performed as follows, based on a published protocol [99].

First, the slides or coverslips (hereafter referred to as “slides”) are placed in a slide staining jar (Coplin-type) and sonicated for 10 min in 1M KOH. The KOH is removed, and the slides are washed at least three times with deionized water. Next, the slides are immersed for 20 min in an aqueous “base piranha” solution consisting of 14.3% v/v ammonium hydroxide and 14.3% v/v hydrogen peroxide that is heated to 60-70 °C. The slides are rinsed at least three times with deionized water (optionally, if fused silica slides are being re-used, they may be heated for ~1 min with a propane torch at this step to burn off any residual microscopic contaminants). The slides are then rinsed once with acetone (HPLC purity or higher).

Next, the slides are immersed in a 2% v/v solution of (3-aminopropyl) triethoxysilane (ATPES) in acetone for 10 min, sonicated for 1 min, and incubated for another 10 min. The APTES/acetone solution is discarded and the slides are immediately rinsed 3-5 times with deionized water, then dried completely under nitrogen flow. The slides are now functionalized with surface amines for further reaction with *N*-hydroxysuccinimidyl esters of polyethylene glycol (PEG) and biotin-PEG.

To functionalize the slides with biotin-PEG and PEG, a 1:10 mixture of biotin-PEG-succinimidyl valerate (biotin-PEG-SVA, MW ~5000, Laysan Bio, Inc.) and methoxy-

PEG-succinimidylvalerate (mPEG-SVA, MW ~5000, Laysan Bio, Inc.) is dissolved in freshly prepared 0.1 M $NaHCO_3$ to a final total PEG concentration of 21.6% w/v. The mixture is briefly centrifuged (1 min at 10,000 rpm in a benchtop Eppendorf microcentrifuge) to remove any suspended air bubbles, and 70-80 µL of the PEG solution is immediately sandwiched between two slides, making sure to exclude air bubbles. The slide sandwiches are kept in a humidified environment in the dark at room temperature for 2-3 h. The slides are then carefully disassembled, placed in a slide staining jar (keeping track of the orientation of the coated side) and rinsed at least three times with deionized water, then dried completely under nitrogen flow.

Remaining surface amines are quenched with disulfosuccinimidyltartrate (sulfo-DST, Soltec Ventures) to reduce nonspecific binding of nucleic acids to the surface, as follows. A 10-mg portion of sulfo-DST is dissolved in 350 µL of 1 M aqueous $NaHCO_3$, briefly centrifuged (1 min at 10,000 rpm in a benchtop Eppendorf microcentrifuge), and 70-80 µL of the solution is immediately sandwiched between two slides with the PEG-functionalized surfaces pointing inward towards the sulfo-DST solution. The slide sandwiches are incubated in a humidified chamber for 30 min at room temperature, then rinsed thoroughly with deionized water and dried completely with nitrogen. The slides are stored in the dark under air for up to 2 weeks, or in a desiccator (preferably under inert gas or vacuum) for several weeks.

3.2.4.2 Sample cells

For prism-type TIRF microscopy experiments, fluidic sample cells are constructed using two pieces of double-sided tape sandwiched between a quartz slide and glass coverslip as previously described [100] (Figure 3.2e). Optional drilling of holes in the backing slide and attachment of Tygon tubing permits convenient buffer exchange, while use of quartz microscope slides permits them to be cleaned with detergent and re-used [100], though cheaper borosilicate glass slides may also be used. After use, these slides can be disassembled and re-cleaned as follows: immerse in boiling water for 30 min; carefully peel off any tape and adhesive with a razor blade; rub slide thoroughly with a thick paste of an abrasive detergent such as Alconox; then rinse thoroughly with deionized water and subject to the cleaning protocol in section 3.2.4.1. Note that no visible residue of adhesive should remain on the slide prior to beginning the protocol of section 3.2.4.1.

For objective-type TIRF microscopy measurements, sample cells are constructed by fixing a cut 1-cm length of a pipet tip (e.g., Eppendorf brand) to a coverslip using epoxy adhesive (Double Bubble, Hardman Adhesives; Figure 3.2c). We have also successfully employed 3D-printed sample cells (Figure 3.2d) that have a smaller area of contact with the coverslip (~0.2 mm^2) and a tapered base that permits the use of as little as 5-10 µL of analyte solution without sacrificing sensitivity. The custom design was prepared in Autodesk Fusion 360 and printed on a ProJet 3500 using the M3 Crystal resin at the highest print resolution of 16 µm per layer. As with the pipet tip sample cells, the 3D-printed sample cells are attached to coverslips with epoxy adhesive, but in this case the attachment is performed with the aid of an electronics vise (e.g., PanaVise) to firmly hold the 3D-printed wells against the coverslip during the application of epoxy to prevent the adhesive from seeping in and clogging the small aperture between the interior of the sample well and the coverslip. While the sandwich-type flow cell can be used on objective TIRF as well, the sample cells constructed from pipet tips or tall 3D-printed wells provide higher sensitivity because of a higher ratio between the volume of analyte solution and the contact area with the coverslip; that is, a larger fraction of the analyte may be captured in a small region of the imaging surface, yielding more detectable molecules per field of view. One drawback of the sample cells constructed from pipet tips and 3D-printed wells is they are both for one-time use only. Regardless of type, the completed sample cells may be stored in a dry, inert, dark environment for several weeks prior to use in SiMREPS.

3.2.5 Surface capture of the target analyte

The following protocol applies to all sample cell types with biotin-PEG-functionalized surfaces, but for the sake of clarity all solution volumes apply specifically to sample wells constructed from cut pipet tips, which were used to collect all data presented in this study. Before imaging, the slide surface is briefly washed with 100 μL T50 buffer (10 mM Tris-HCl, 50 mM NaCl, pH 8.0) followed by the addition of 40 μL of 1 mg/ml streptavidin to the sample well. After 10 min, the streptavidin solution is removed and the surface is washed three times with 100 μL of 1× PBS. The surface is then incubated with 40 μL of a solution containing 100 nM of the appropriate biotinylated LNA capture probe in 1× PBS buffer for 10 min. The solution is removed and then the sample cell is washed three times with 100 μL of 1× PBS. Finally, a 100-μl portion of sample containing the target RNA or DNA and 2 μM carrier oligonucleotide is introduced into the sample chamber and incubated for 1 h to capture the analyte at the imaging surface. Note that double-stranded DNA samples must first be denatured by, for example, heating to 95 °C in the presence of 2 μM carrier oligonucleotide, then cooling to room temperature in a water bath for 5 min before adding to the sample cell. For direct capture of analytes from crude biofluids such as cell extract or serum, a pre-incubation step in ~2% (w/v) sodium dodecyl sulfate (SDS) and 0.16 U/µL of proteinase K (New England BioLabs, Inc.) is used to liberate nucleic acids from any protein binding partners as well as to inactivate any nucleases present in the sample [86]. After the 1-h capture incubation, the sample solution is removed, and 1× PBS buffer is added to the sample cell until the imaging buffer (see section 2.6) is added. Note that, while analytes can be captured from crude biofluids [86], the imaging should still be performed in a standard imaging buffer to ensure reproducible probe binding and dissociation kinetics.

3.2.6 Imaging

All data discussed in this paper were collected using an Olympus IX-81 objective-type TIRF microscope equipped with a 60× oil-immersion objective (APON 60×OTIRF, 1.49 NA) with both Cell^TIRF and z-drift control (ZDC2) modules, and an EMCCD camera (IXon 897, Andor, EM gain 300). Cy5 excitation was provided by a 640-nm red laser (Coherent CUBE 640-100C, 100 mW) and Cy3 excitation by a 532-nm green laser (CrystaLaser CL532-150mW-L). To delay the photobleaching of fluorophores and thus obtain more accurate measurements of the bound-state lifetime of the FP, a 25 nM solution of the FP is prepared in an imaging buffer containing 4× PBS, 2.5 mM 3,4-dihydroxybenzoate, 25 nM protocatechuate dioxygenase, 1 mM Trolox (oxygen scavenger system, OSS [101]), and added to the sample chamber for SiMREPS imaging. The imaging solution for *EGFR* L858R mutant and wild-type discrimination in this study also includes 100 nM of a WT competitor sequence to block FP binding to the WT *EGFR* sequence. Usually 3-5 minutes are allowed for the OSS to achieve a low steady-state oxygen concentration before imaging. The transient binding of FP to captured target molecules is monitored for 10 min under TIRF illumination, with a movie acquisition rate of 2 Hz and an EM gain setting of 150. All imaging is performed at a darkened room at an environmentally controlled temperature of 20 ± 3 °C.

The high ionic strength of the imaging buffer promotes rapid binding of the FP to the target [96], allowing for many cycles of FP binding and dissociation within the 10-min observation period for well-optimized FP sequences. The concentration of FP in the imaging buffer may be adjusted, but typically is optimal in the range of 25-50 nM; lower concentrations will reduce the frequency of FP binding, while much higher concentrations will result in prohibitively high levels of background fluorescence from freely diffusing probes during imaging. If dissociation kinetics of the FP are relatively slow, for instance due to a longer or more GC-rich FP sequence, denaturants such as 10-30% formamide can be used to decrease the duration of the bound state, albeit at greater risk of target dissociating from the CP during the experiment.

The length of the observation period for each field of view is a particularly important parameter, since enough time must be allowed for multiple (e.g., >10) cycles of binding and dissociation to each surface-bound analyte molecule, thus permitting adequate separation between specific and nonspecific binding signatures for zero-background measurements. The exact imaging time required is dependent on the kinetics of specific and nonspecific binding, as well as the degree of separation between signal and background peaks that is desired. A useful guideline for selecting a minimum observation time is embodied in the following relationship [86] :

$$t \ \geq \ 2s^2 \frac{k'_{bind}+k_{diss}}{k'_{bind}k_{diss}} \frac{\left(1+\sqrt{f}\right)^2}{(1-f)^2}, \tag{1}$$

where *t* is the observation time, *s* is the desired number of standard deviations separating the signal and background peaks, k'_{bind} is the pseudo-first order binding rate constant for the query probe to the target, k_{diss} is the first-order dissociation rate constant of the query probe from the target, and $f = \frac{\langle N_{b+d}\rangle_{nonspecific}}{\langle N_{b+d}\rangle_{specific}}$ is the ratio between the average number of nonspecific binding and dissociation events observed per trace ($\langle N_{b+d}\rangle_{nonspecific}$) and the average number of specific binding and dissociation events observed per trace ($\langle N_{b+d}\rangle_{specific}$). For example, if a separation of *s* = 3 standard deviations is desired between signal and background, and with k'_{bind} = 5 min^{-1}, k_{diss} = 5 min^{-1}, and $f = \frac{\langle N_{b+d}\rangle_{nonspecific}}{\langle N_{b+d}\rangle_{specific}}$ = 0.1, the minimum observation time is 3.9 min. The sampling interval (exposure time per frame) should be significantly less than the smaller of τ_{bound} and $\tau_{unbound}$; in the above example, significantly less than 0.2 min, e.g., ~1 s per frame (sampling frequency of ~1 Hz).

3.2.7 Data analysis for kinetic fingerprinting

All MATLAB scripts for SiMREPS data analysis are in a publicly available github repository (https://github.com/ajohnsonbuck/simreps-2018-08). A typical analysis of movies from SiMREPS experiments consists of the following steps: 1) identification of "candidate" regions of interest (ROIs) within the image exhibiting greater frame-to-frame intensity fluctuations than their surrounding pixels (Figure 3.3a, b); 2) calculating the frame-by-frame fluorescence intensity of each ROI (Figure 3.3c), 3) hidden Markov modelling (HMM) to calculate FP binding and dissociation kinetics for single-molecule kinetic fingerprinting (Figure 3.3d); and 4) application of filters to distinguish nonspecific

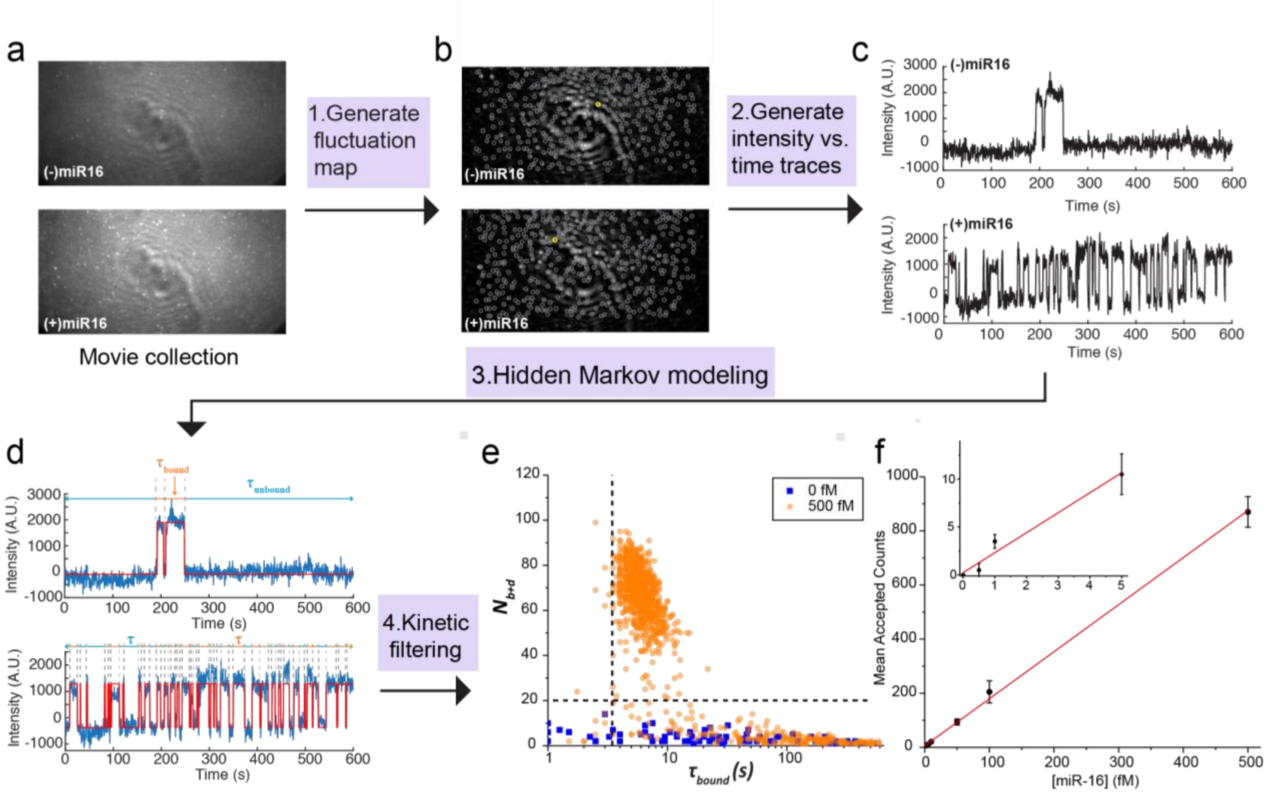


Figure 3.3. Data analysis pipeline. **a**, Single-frame images of representative fields of view from TIRF microscopy. **b**, Intensity fluctuation maps of the fields of view shown in **a**. Grey circles indicate positions of local maxima in the fluctuation map, from which candidate ROIs are identified for further analysis by generation of intensity vs. time traces. **c**, Representative intensity vs. time traces generated from the ROIs identified in **b**, circled in yellow. **d**, HMM idealization (red lines) for each intensity vs. time trace. Bound and unbound-state dwell times (τ_{bound} and $\tau_{unbound}$, respectively) are indicated by the orange and blue horizontal line segments above the idealization. **e**, Candidates in the positive (orange circles) and negative (blue squares) controls for miR-16 are well separated by thresholds of N_{b+d} > 20 and τ_{bound} > 2.5 s (black dashed lines), permitting discrimination of specific and nonspecific binding at the single-molecule level. Data are pre-filtered for signal-to-noise > 2.5 and intensity > 1000. **f**, miR-16 standard curve. n = 3 replicates for blank, 2 replicates for other measurements. Error bars represent 1 standard deviation.

from specific binding based on signal-to-noise, intensity, and FP kinetics (Figure 3.3e). Prior to step 1), a software-based drift correction may be applied to compensate for lateral stage drift during the experiment, though this is often not necessary if the microscope system is sufficiently stable (e.g., < 3 pixels of drift during the 10-min movie). candidate region. Upon request, MATLAB scripts for all the necessary processing steps below can be provided.

3.2.7.1 Identifying candidate ROIs

For optional drift correction, a custom routine written in Matlab (available upon request) based on the subpixel correlation between consecutive recorded images can be used to compensate for any x-y stage drift that would interfere with subsequent intensity-versus-time analysis of candidate ROIs. After this optional step, candidate ROIs—generally 5-pixel×5-pixel regions with significant frame-to-frame intensity fluctuations—are identified as follows. Each of the *N* movie frames is subtracted from the previous frame and the absolute value taken to generate a new image of the same dimensions as the original, but in which each pixel value represents the absolute value of the intensity change from the previous frame to the current frame. This is repeated for all movie frames, resulting in a new image stack with (*N*-1) frames. Finally, the value of each pixel in this image stack is averaged, resulting in a single image called a "fluctuation map" containing the average frame-to-frame change in intensity for each pixel. Pixels representing local maxima within this image are selected to serve as the center pixel of each candidate ROI for further processing.

3.2.7.2 Calculation of intensity-*versus*-time traces

The intensity-*versus*-time trace for each candidate ROI identified from the fluctuation map is generated as follows. Within the first frame of the *original* movie file, the intensity of all 25 pixels within the 5-pixel-×5-pixel ROI is summed to create a single fluorescence value, and the median intensity value of the 2-pixel-wide region surrounding the ROI is subtracted to find the background-subtracted intensity of this ROI within the first frame. This process is repeated for each frame of the movie, and the list of intensity values combined to create an intensity-*versus*-time trajectory for this ROI. The process is repeated for each ROI identified from the fluctuation map, and the intensity-*versus*-time trajectories are exported as an ASCII file for import into the HMM software QuB [102].

3.2.7.3 Hidden Markov modeling

The traces are imported into the HMM software QuB and fit using a two-state model. Proper parametrization is essential for convergence of HMM fitting; that is, the amplitudes, standard deviations, and kinetics should be as close as possible to the expected behavior of the FP binding to the target, and ideally within ~1 order of magnitude. It is important to use the same model to fit all datasets that are to be compared. The HMM fitting results table from QuB is exported for further analysis of the intensity and kinetics in MATLAB.

3.2.7.4 Filtering specific from nonspecific binding

A binary classification is performed on each candidate ROI based on whether its intensity-*versus*-time trace satisfies certain criteria. The criteria are established by an empirical evaluation of traces collected in negative and positive control experiments—e.g., in the absence and presence of 500 fM synthetic target nucleic acid—and chosen so as to reject essentially all traces in the negative controls while accepting as many traces as possible in the positive controls. Since nonspecific binding of the probe to the surface can vary somewhat between coverslip or slide preparations, it is generally advisable to establish these criteria based on several independent technical replicates, preferably on different days. While the specific criteria will vary depending on factors such as the target, FP sequence and concentration, imaging buffer, and acquisition temperature, in this study a candidate ROI is considered to contain a true positive signature of the analyte if it satisfies the following criteria:

- Intensity difference between bound state and unbound state (ΔI) > 1000 counts for detection of miR-16, > 500 counts for detection of *EGFR* L858R
- Signal-to-noise ($\Delta I/\sigma$, where σ is the standard deviation of the intensity in the FP-unbound state) > 2.5 for miR-16, > 2 for *EGFR* L858R
- Number of FP binding and dissociation events per observation period, $N_{b+d} \geq 20$
- Median lifetime in the FP-bound state, $\tau_{bound,median}$ > 4 s for miR-16, > 5 s and < 20 s for *EGFR* L858R
- Median lifetime in the FP-unbound state, $\tau_{unbound,median}$ > 0 for miR-16, > 20 s and < 50 s for *EGFR* L858R

All traces satisfying these criteria are counted as true positives, and those that do not are considered to show insufficient evidence to be counted as true positives. Of the above criteria, the most critical for rejecting false positives (as determined from negative control measurements) tend to be N_{b+d} and $\tau_{bound,median}$.

3.3 Results

3.3.1 Considerations for the use of SiMREPS to detect short nucleic acid

In negative control measurements with imaging buffer containing the FP, but in the absence of the target analyte, a considerable number of FP binding events were always observed—typically numbering in the hundreds—suggesting that transient or long-lasting interactions between the FP and the imaging surface were difficult to suppress entirely (Figure 3.1d). In a conventional analysis without kinetic fingerprinting, it would be necessary to subtract these counts from all measurements as background; however, the large standard deviation of this background (Figure 3.1d) would impose a limit of detection (LOD) of hundreds of captured target molecules per FOV.

In contrast, by applying the kinetic filtering criteria as outlined in section 2.7, essentially all of these background counts were filtered out in the negative control experiments (Figure 3.1d), permitting the confident identification and counting of even single-digit numbers of target molecules per FOV. This is because, through repeated sampling of the same surface-immobilized target molecules through multiple cycles of FP binding, a progressively better estimate of kinetic parameters such as N_{b+d}, $\tau_{bound,median}$, and $\tau_{unbound,median}$ was obtained for each candidate ROI, and it became easier to resolve true and false positives by a binary classification based on the kinetic criteria outlined in section 2.7 (Figure 3.3e). The number of accepted counts (candidate ROIs that pass kinetic filtering) was linear within the range of approximately 1-800 molecules per FOV, as shown by the standard curve for miR-16 (Figure 3.3f). Due to the essentially zero background, even 0.5 fM miR-16 yielded significant counts above the negative control, resulting in an LOD that was mainly limited by the capture efficiency of analyte on the imaging surface rather than on background binding of the FP or autofluorescence of the imaging surface. In terms of absolute concentration [103], the calculated limit of blank (LOB) of this assay is 0 (since no blank counts were detected), and the estimated LOD is 0.4 fM.

If more than ~500 molecules are present in a FOV, the diffraction-limited analysis presented here will result in a sub-linear increase and eventually a decrease in the accepted counts due to the inability to resolve closely spaced molecules. If it is desired to extend the dynamic range beyond this ~2.5 orders of magnitude into the range of thousands of molecules per FOV or more, it will likely be necessary to switch to a more conventional quantification scheme based on fluorescence intensity, or to implement super-resolution methods to analyze the kinetics of FP binding with sub-pixel accuracy [104]. Indeed, one recent paper describes the use of super-resolution imaging and kinetic analysis of dissociation kinetics to discriminate single-nucleotide variants in DNA with 95% accuracy [105].

3.3.2 Highly selective detection of *EGFR* L858R

We tested the ability of SiMREPS to distinguish between closely related sequences, using as a model of the point mutation *EGFR* L858R (c.2573T>G), a common driver mutation in non-small cell lung carcinoma. Note that the high GC content surrounding this mutation necessitated two design choices for the FP: the intentional introduction of a G-T wobble mismatch in the FP-target interaction, and the positioning of the mutation towards one end of the FP-target duplex to reduce the GC content slightly (Figure 3.4a). While considerable FP binding was observed in the presence of the wild-type (WT) sequence, the traces in the WT-only experiment could be distinguished from the mutant

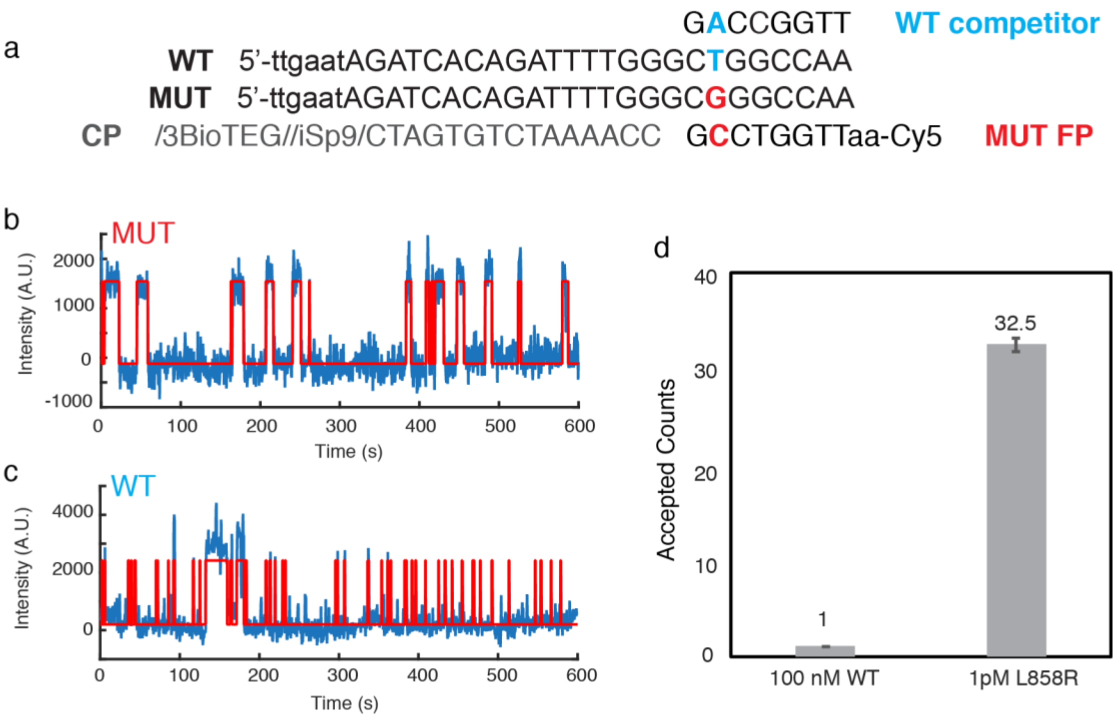


Figure 3.4 Single-base selectivity of SiMREPS. **a**, Sequences of WT and L858R MUT targets, as well as the capture probe (CP), MUT fluorescent probe (FP) and WT competitor. **b**, Representative intensity vs. time trace from MUT-only positive control. **c**, Representative intensity vs. time trace from WT-only control. **d**, The accepted counts after kinetic filtering of traces collected in the presence of 100 nM *EGFR* WT or 1 pM L858R MUT. The apparent discrimination factor between MUT and WT is 3.25 million.

(MUT) traces on the basis of the median bound-state lifetime ($\tau_{bound,median}$), which was longer for some traces in the presence of the MUT (Figure 3.4b,c). Indeed, the number of accepted traces in the presence of 100 nM WT was > 30-fold lower than in the presence of only 1 pM MUT, despite the fact that the WT was present at a 100,000× higher concentration. The apparent discrimination factor of this assay is thus approximately 100,000 x 32.5 / 1, or 3.25 million (Figure 3.4d). This is far greater than the theoretical maximum for thermodynamic binding assays of any point mutation ($Q_{max,therm}$), and demonstrates the power of SiMREPS to discriminate between very closely related analytes, entirely without amplification

3.3.3 DNA CpG methylation detection using kinetic fingerprinting

SiMREPS has been proved to be a powerful detection technique for short nucleic acids such as miRNA and ctDNA[31,86]. The current design is able to distinguish a single nucleotide difference between two analyts such as miRNA *let-7a* and *let-7c*. In other words, SiMREPS is sensitive enough to detect the kinetic binding difference from one single nucleotide. A recent work from Walter's group[31] had an interesting observation when the researchers found out that the deaminated wild type target caused the false positive signals in T790 wild type detection. The true positive signal should come from mutant T790 (C → T mutation). From their observation, the target that has deaminated WT, in which case one cytosine (C) in the sequence becomes a deoxyuracil (dU), gave similar signal as the mutant group. But the histogram of N_{b+d} between mutant group and Deaminated group are not the same (Figure 3.5). A slight separation of the histogram can be seen. We suppose a more obvious separation should be seen if given a longer observation time (currently 10 minutes).

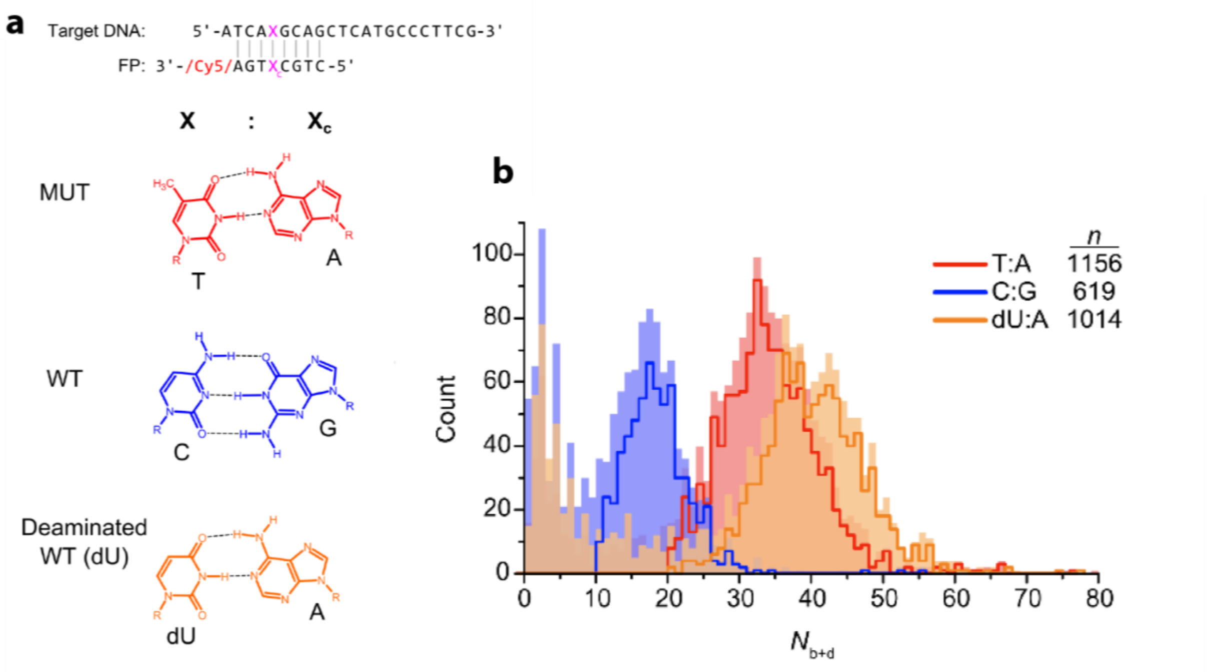


Figure 3. 5. SiMREPS detection results of wild type T790 (WT), mutant T790 (MUT) and Deaminated wild type T790 (dU). **a**, Diagram showing the various Target DNA-FP combinations tested, as well as expected base-pairing interactions. The 3' barcode sequence (TAGGAC) present on the S17 Target DNA is omitted for clarity. dU, deoxyuracil. **b,** Histogram of Nb+d for different combinations of FP and target DNA per 10-minute experimental observation. The shaded regions indicate the Nb+d distribution for all molecule candidates prior to kinetic filtering, and solid lines represent the distribution of candidates that pass filtering (i.e., apparently genuine target molecules). n, number of apparently genuine target molecules that pass filtering. Reprinted from HaywardLund_et_al_JACS2018SI. Copyright 2018.

Since the separation of N_{b+d} histogram is an important metric in SiMREPS detection, this observation showed that the binding kinetics of thymine and deoxyuracil to the same FP are different. The only difference here in this two cases is the deoxyracil has an extra methyl group than thymine. This result opened the possibility of using SiMREPS to detect the methylated DNA and unmethylated DNA. The most important DNA methylation is the CpG sites.

DNA methylation at CpG sites is associated with genomic imprinting, X-chromosome inactivation and the mechanism of carcinogenesis across diverse cancer types[106–108]. The detection of methylated DNA sequence in biofluids has shown great potential for non-invasive of cancer[109,110]. For instance, the CpG methylation at the vimentin and septin9 DNA promoter loci occurs early during the process of carcinogenesis and this DNA is detectable in patient blood plasma samples. However, measurement of DNA CpG methylation with high specificity and sensitivity has been very challenging. This is mostly due to the reason that methylation marks cannot be directly PCR amplified in the same way as mutations can, either as being read directly by sequencing methods. Currently, the gold standard assay is using bisulfite treatment to convert 5mC to uracil, followed by DNA sequencing using primers sensitive to the conversion. The problem of this method is that the bisulfite treatment is typically damaging or destroying 80% of the DNA and thus limiting the sensitivity and specificity of this method. As a single molecule technique, SiMREPS can meet the requirement of such high sensitivity and specificity without the need of bisulfite treatment. We detected the wild type T790(WT) and methylated T790(5mC) with same FP (Figure 3.6). The observation time was 30 minutes to get a better separation of the histogram of N_{b+d}.

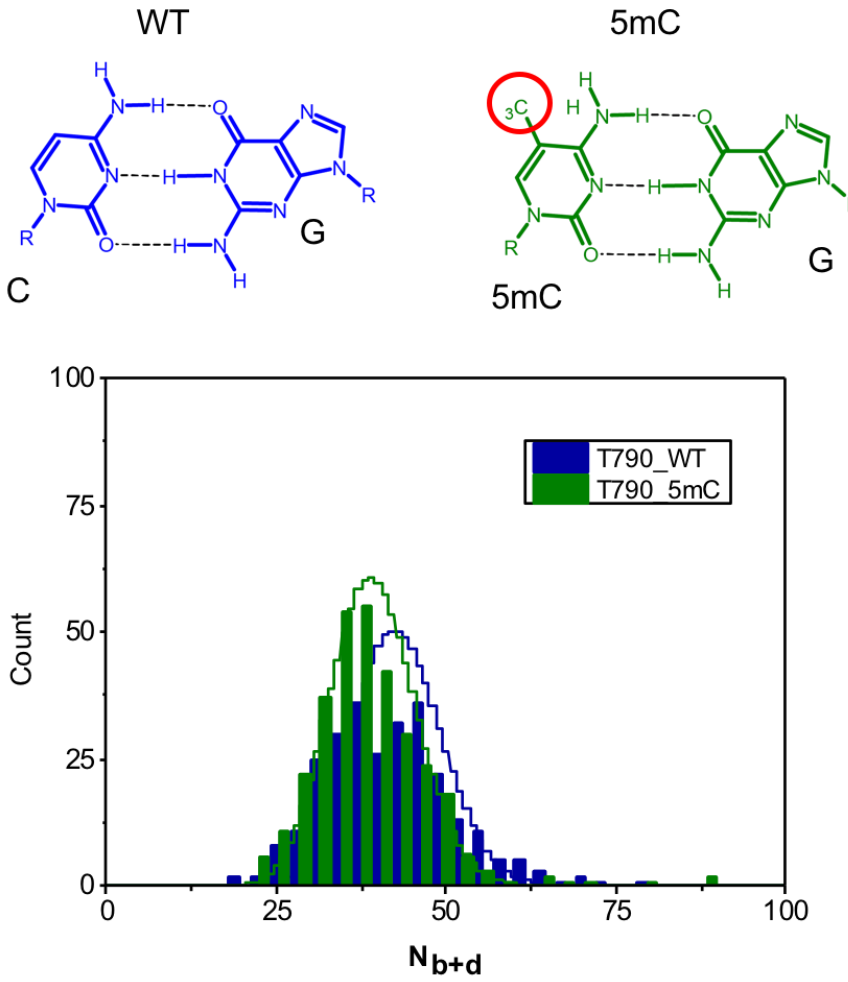


Figure 3. 6. Histogram of N_{b+d} for Methylated T790 (T790_5mC) and wild type T790 (T790_WT). No clear separation of the number of bound and unbound states between the wide type and methylated T790 under the imaging time of 30 minutes.

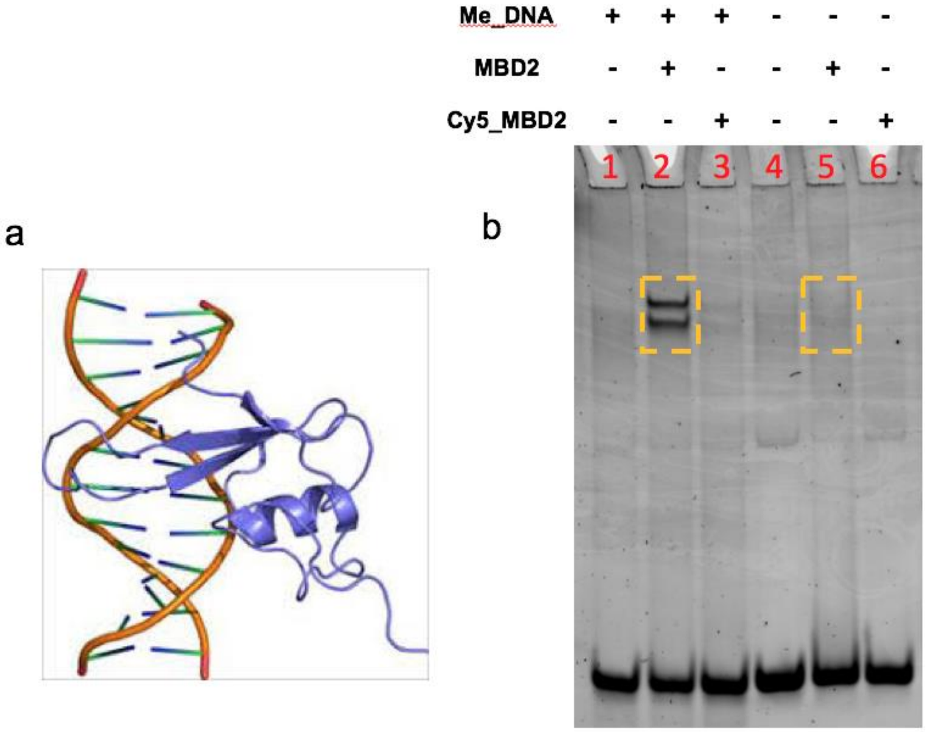


Figure 3. 7. **a**, schematic of MBD binding to methylated DNA double helix. **b**, Native PAGE results of MBD2 selectively binding to methylated and unmethylated DNA.

However, as the histogram shows, the separation between wild type and methylated T790 is not as good as expected, even for 30 minutes' observation time. One possible reason is that in GC pairs the impact to binding kinetics from one methyl group is not as strong as in AT pairs because GC pairs form 3 hydrogen bonds while as AT pairs form 2 hydrogen bonds.

Since SiMREPS is a kinetic fingerprinting technique, any transient binding process should be detectable in theory. Bindings such as nucleic acids interaction, protein-protein interaction, protein-nucleic acid interaction should all be detectable. We switched the fluorescence probe from short oligonucleotide to fluorophore labeled small protein Methyl-CpG Binding Domain (MBD). The MBDs are a family of proteins that show higher affinity for methylated DNA than for unmethylated DNA as in the nanomolar range than micromolar range. The Kd for MBD2 binding to methylated DNA is ~3nM, while the Kd for MBD2 binding to unmethylated DNA is > 200nM[108]. The non-denaturing PAGE result (Figure 3.7) shows that the MBD2 has a selectivity over methylated DNA and unmethylated DNA at an ensemble level. After successfully labeling the fluorophore, we think the MBD2 will be a good FP candidate for DNA CpG methylation detection.

3.4 Discussion

We here have presented a workflow for the detection of nucleic acid targets by single-molecule kinetic fingerprinting through SiMREPS, and shown that this method affords detection of single analyte molecules with essentially no background (0-1 counts per FOV) in negative controls, even when challenged with a large concentration of closely related sequence. We further show that the single-base selectivity of the technique is sufficient to detect a mutation as subtle as a single T-to-G substitution with an apparent discrimination factor > 1 million, far in excess of any other amplification-free technique and comparable to the best available methods (*i.e.*, droplet digital PCR). The ability of SiMREPS to accommodate very short (< 25 nt) analyte sequences, and those captured from crude biofluids with minimal processing, are unique advantages relative to most amplification-based methods. To make the technique more widely applicable and convenient, future improvements may include the use of techniques to improve mass transfer of analytes to the surface in order to increase the density of captured analyte, thus increasing sensitivity; modified probe or assay designs to permit more rapid cycling between bound and unbound states to shorten the imaging time needed to reach any desired level of specificity; and/or automated signal detection and counting algorithms. For instance, while published data here and elsewhere [86] indicate typical limits of detection of ~1 fM for passive analyte capture in our standard pipet-tip sample cells (Fig. 2c), further exploratory work suggests that attomolar detection limits may be achievable in the near future (data not shown); furthermore, in theory, even single-digit copy numbers could be detected with sufficiently high capture efficiency. Furthermore, there is no fundamental limit to the type of analyte that can accurately be detected and quantified using SiMREPS, making it a universal platform that – with further refinements – may transform biomarker detection just as super-resolution has conventional fluorescence microscopy

Chapter 4

Automatic Traces Selection Using Machine Learning[5, 6]

4.1 Introduction

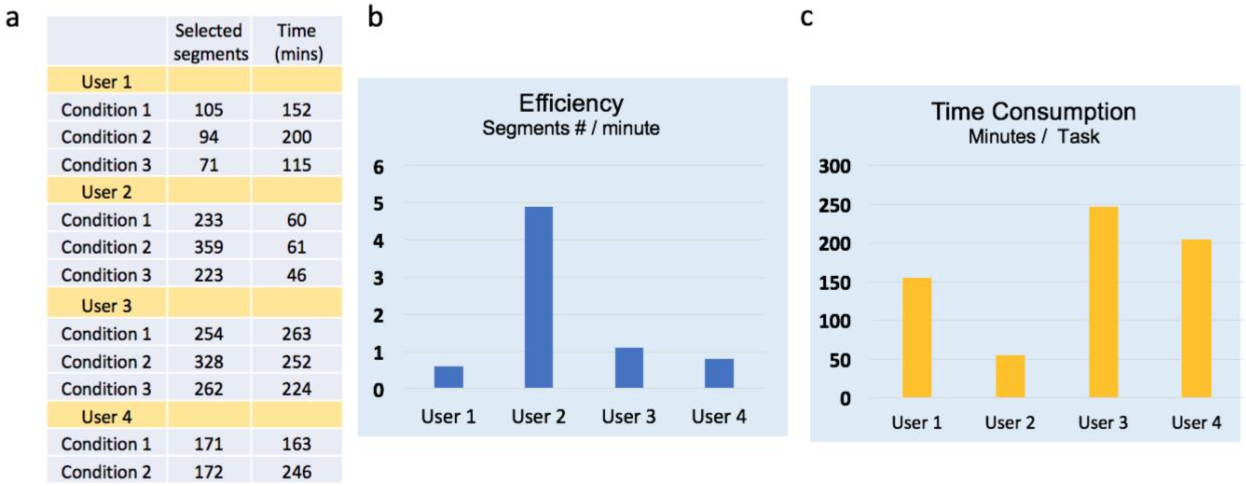
Single-molecule fluorescence microscopy has been a powerful technique enabling investigation for structural dynamics of biomolecules, especially when ensemble averaging or lack of synchronization might shadow the detailed information of the system under study. One leading research direction of this field is developing new imaging methods to achieve higher spatial resolutions such as DNA-PAINT, STORM, etc.[111–113] Besides the experiment techniques exploration, data analysis optimization takes a significant role in single-molecule fluorescence microscopy study. For instance, Ha's group established a now regularly used trajectories analysis method for single-molecule fluorescence or Förster resonance energy transfer (FRET) data using hidden Markov Modelling (HMM)[114]. Walter's group reported a hierarchical clustering of hidden Markov modeling-fitted smFRET trajectories enabling rapid interpretation of complex single-molecule behaviors[115].

[5] Jieming Li and Leyou Zhang initiated the idea. Leyou Zhang designed and implemented machine learning algorithms. Jieming Li collected training and test datasets. Jieming Li performed the validation of machine learning algorithm analysis results. Jieming Li, Leyou Zhang and Alexander Johnson-Buck analysed and interpreted the data.

[6] This chapter will be rewritten into a manuscript for publication submission.

Through the pipeline of analyzing single-molecule measurement, sorting and classification of single-molecule time series data (such as smFRET data) is a critical but time-consuming part. There has been no automatic method that can select out the qualified single-molecule trajectories. This is because the broad diversity of potential background signals makes it difficult to design simple criteria, such as thresholds based on intensity or noise, that would effectively remove all irrelevant time traces (e.g., those resulting from contaminants or nonspecific binding rather than analyte molecules) while retaining all or most of the relevant data for further analysis. Furthermore, the heterogeneous dynamic behaviors of molecules from experiment to experiment may require different selection criteria such as the degree/level of anti-correlation of signals. In addition, it is often the case that only a particular time segment of each single-molecule trace is useful, and these regions of interest (ROIs) must typically be selected by hand due to this reason, which slows down analysis considerably. Because the selection of ROIs is dependent on relatively complicated determinations – such as the absence of various artifacts including photobleaching, blinking, contamination from nearby fluorescent materials, inconsistency of the segment duration – it is also difficult to automate this process using conventional methods.

In most cases, hundreds of hours are spent on manual selection for an operator before they can get an adequate number of qualified molecules for further study. A user's behavior study investigation was done to researchers anomalously in Nils Walter Lab. (Figure 4.1) We calculated the number of saved trace segments and the corresponding time spent on that selection (Figure 4.1a). Among the four users randomly picked, the fastest user can select ~5 segments per minute, while the slowest user only selects less than 1 segment per minute. If we count by the task (here as one dataset from one experiment condition’s measurement), even the “fastest picker” needs ~1 hour for one task. The longest time here is more than four hours for one task. If the



	Selected segments	Time (mins)
User 1		
Condition 1	105	152
Condition 2	94	200
Condition 3	71	115
User 2		
Condition 1	233	60
Condition 2	359	61
Condition 3	223	46
User 3		
Condition 1	254	263
Condition 2	328	252
Condition 3	262	224
User 4		
Condition 1	171	163
Condition 2	172	246

Figure 4.1. Users behavior investigation of 4 researchers. **a,** Table of segments number and time spent on selecting the corresponding segments. The calculation is by accumulation of time intervals from the time record of the auto-saved segment files on data server. The dataset analyzed was randomly picked. The only criterion was to include as many segments as possible from one condition. **b**, Average number of segments can be selected by different users in a time unit as one minute. **c**, Average time needed by different users to finish one task, here as manual selection of data from one experiment condition.

data is more complex, such as molecules having more heterogeneous behaviors, the time spent on the human selection may be even longer. Improving the productivity of data analysis in smFRET is highly demanding.

In addition to sorting smFRET traces, improving the selectivity of kinetic fingerprinting is another incentive. As discussed in Chapter 3, the current data analysis in SiMREPS kinetic fingerprinting requires Hidden Markov Modeling (HMM) fitting to recognize fluorescence intensity change. After HMM fitting, a kinetic thresholding is applied on several parameters such as minimum binding time (τ_{bound}) and the number of transitions (N_{b+d}) to screen out false positive traces from non-specific binding. However, this method, which is referred as HMM Thresholding in later discussion, is not perfect: 1) The HMM fitting software QuB does not work well all the time and sometimes fails at a fitting. This rare mistake might happen to a few traces among hundreds of trances, so it is within the error tolerance for the general fitting purpose. However, for low concentration detection where only the single digital number of molecules can be detected this fitting failure reduces the specificity a lot; 2) The criteria of kinetic thresholding are established by empirical evaluation of traces collected in negative and positive control experiments. Thus, the kinetic thresholding will vary depending on factors such as FP sequence and concentration, imaging buffer, and acquisition temperature. The empirical evaluation itself might also be impacted by subjectivity from person to person. For example, the dashed line as indicating minimum N_{b+d} can be slightly moved to get more accepted molecule numbers or fewer (Figure 3.3e). In other words, the molecules at the kinetic threshold edge are hard to classify as accept or reject; 3) The current kinetic thresholding uses nine parameters to screen traces. They are intensity difference between bound state and unbound state, signal to noise, number of binding and dissociation events, median lifetime in bound and unbound states, minimum lifetime in bound and unbound states, and maximum lifetime in bound and unbound states. Even though the current parameters work well to accept as many molecules as possible in positive groups while achieve a zero background in negative groups, we believe more parameters can be used to improve the performance of traces classification, thus improving the specificity of SiMREPS in low concentration detection.

With the explosion of “Big Data” problems, applications of machine learning in life science researches are growing. Xiong H. Y. et al. developed a machine learning technique that scores how strongly genetic variants affect RNA splicing to facilitate precision medicine and whole-genome annotation[48]. Christiansen E.M. et al. established a computational machine learning approach as they call “in silico labeling” to predict some fluorescent labels from transmitted-light images of unlabeled fixed or live biological samples[116]. The in silico labeling can generate biological measurements that would be otherwise problematic or impossible to acquire. A study more related to super-resolution imaging is from Ouyang W et al.[117] In this work, they presented a computational strategy that uses deep learning neural networks to reconstruct super-resolution views from sparse, rapidly acquired localization images and wide field images that would otherwise require 10 times more images acquisition. All those inspiring studies have broadened the application spectrum of machine learning in life science and therefore makes machine learning become a more accessible and trustworthy approach in data analysis.

The task of sorting time series data generated from single molecule measurement is very similar to a clustering problem in machine learning. A similar approach has been applied in nanopore sequencing of DNA[118]. In this study, the data being analyzed is single channel time series data from the noisy and complex electrical signals. The authors trained a combined neural network that has both convolutional neural network (CNN) layers and recurrent neural network (RNN) layers. In this way, the new network can translate raw electrical signal directly to the nucleotide sequence. Following the idea of analyzing similar-look data, we used RNN and CNN in the analysis of single molecule traces. The RNN network we use here is Long Short-Term Memory (LSTM) network, which is suitable for classifying, processing and making predictions based on time series data[119]. The CNN network we use here is AlexNet[46], which is famous for its capability of getting extremely high accuracy of image recognition with a relatively small number of network layers (8 layers). For single channel data, the automatic selector developed here can accept more traces in positive groups than the HMM Thresholding method, while keeping negative control groups accepting almost zero molecules. For double channel data, the automatic selector's prediction agrees 96.1% with human selection as the highest output. We hope this automated selector can help researchers improve the data analysis productivity in smFRET measurement. In the meanwhile, the machine learning network can be used as a model-free traces classifier in SiMREPS measurement to achieve a higher selectivity.

4.2 Materials and Methods

Data preprocessing is essential in machine learning, which relies heavily on consistent training data. For double-channel data from smFRET measurement, we used a dataset consisting over 20,000 FRET traces by extracting from four persons' experiment data. These four people might have slightly different selection criteria such as traces duration time/segment length and degree of anti-correlation. These data have been screened before to ensure that they meet criteria of total length and intensity. In the dataset, the ratio between human selected traces and unselected traces are close to 1:3. To balance the dataset, we include additional 5,000 simulated traces generated by HMM with kinetic constants and transition matrix set to representative experiment systems. For single channel data from SiMREPS measurement, data collection is straightforward as no human selection label is required for training. We used T790M detection data consisting of data from mutant DNA and wild type DNA experiments for training. The T790M mutation results in an amino acid substitution at position 790 in EGFR. All data were screened by length and intensity criteria as well.

The recursive neural network (RNN) is feed with raw traces data as input. The network uses an input layer, a layer of bidirectional Long-Short Term Memory (LSTM) cells, a fully connected layer, and a SoftMax classification layer. For all the layers, we used the MATLAB(2018a)'s Neural Network package. The input layer consumes a raw trace as a vector of dimension $[1, T]$ and converts it into a matrix of dimension $[n_d, T/n_d]$, by which each n_d nearest data point is binned into one row. We found $n_d = 10$ to produce the best training results. If n_d is too small, the information of each column is insufficient for feature detection in the presence of noise. If n_d is too large, it gives a high dimensional feature space which hinders training. The matrix is feed into the LSTM layer, which contains 200 LSTM cells. Bidirectional LSTM is used because it enables later frames to influence classifying earlier frames, which is a useful ability for the model. For example, some traces are good initially but are bad for the rest of the frames. The bidirectional LSTM model will use all the information, not just the good initial frames, during trace segmentation for good frames. The classification and segmentation is completed by the fully connected layer and the SoftMax classification layer. Together, they take the output of LSTM layer of dimension $[200, T/n_d]$ as input and a classification label as output. If only the classification of trace is needed, only the last row in the input is used as the feature. Otherwise, we use every row for frame-wise classification.

Aside from RNN, we deployed a convolutional neural network (CNN) for FRET traces classification. Raw traces are first converted into 2-D images by plotting them as scatter plots with donor and acceptor intensity as x and y position respectively. The 2-D images represent the intensity pattern of the raw traces but lose all the time information. If the FRET trace represents a good trace with a correlation in donor and acceptor intensity, the 2-D image will have a clear slope = -1 pattern of clusters of scatter points. We down-sampled the 2-D scatter plots to 40×40 in order to reduce feature space. The images are fed into AlexNet, a pretrained 8-layer network, for training. Each image is associated with a label of either human selected or unselected. The AlexNet is easier to converge than a network that was built from scratch since its parameters are initialized during pretraining. For our dataset, we found that training the network around 20 hours on a modern CPU is sufficient. On a GPU, we are able to achieve convergence within 4 hours using NVIDIA's TESLA V100. The scatter plot generation, down-sampling and network training were all implemented with MATLAB (2018a).

4.3 Results

4.3.1 Use machine learning to analyze single channel data from kinetic fingerprinting measurement

As discussed before, the currently used HMM Thresholding is not perfect due to the HMM fitting failure happens every now and then. Rather than using any kinetic binding parameters as criteria to screen out the false positive traces, we label the traces groups as positive and negative based on the experiment conditions, and let the RNN network extract the different features among the traces and make the decision during the training (Figure 4.2). The only criterion is to let the network accept as many traces from positive groups while accepting no traces from negative groups. Here we used the data collected from T790M detection[31]. The mutant groups are labeled as "Positive" and the wild type groups are labeled as "Negative". Obvious transitions can be seen in representative mutant traces. After training, the LSTM network makes a decision on each trace as accept (check mark in Figure 4.2) or reject. In the traces classification step, the trained LSTM can assign a score ranging from 0 to 1 to each input unlabeled trace. "1" as having the highest probability to be a mutant trace while "0" as having the least probability to be a mutant trace. In this dataset case, "0" can also be interpreted as traces having high similarity to traces from wild type groups. A natural threshold of the score is 0.5 since it's a binary classification. So, any traces with a score higher than 0.5 will be taken as "Accept" and any traces with a score equal to or lower than 0.5 will be taken as "Reject".

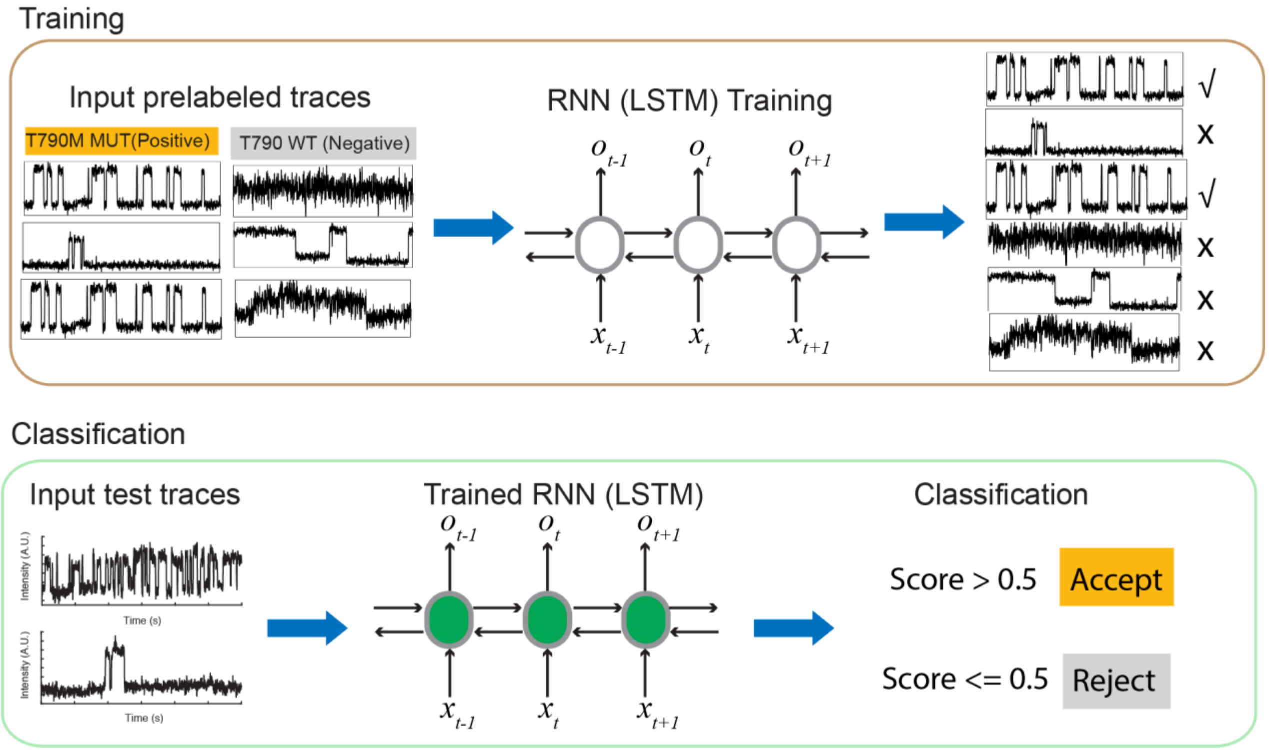


Figure 4. 2. Schematic of training and classification using machine learning algorithm.

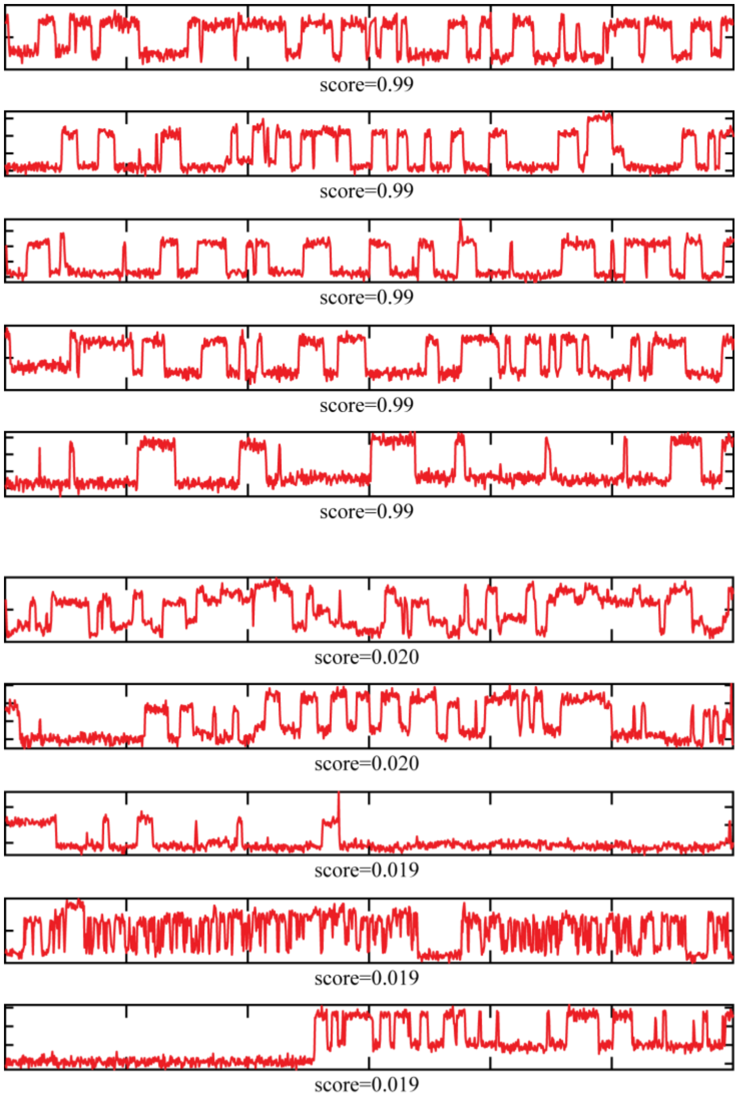


Figure 4. 3. Representative traces with high and low scores assigned by machine learning algorithm.

Representative traces assigned with high scores and low scores are shown in Figure 4.3. High-score traces show clear repetitive transitions between two states and good signal-to-noise. Low-score traces have obvious “unacceptable” features such as more than two states intensity changes, non-repetitive transitions, or fast transitions indicating bindings not from nucleotide hybridization. The key problem in this sorting task is whether the network can recognize “bad features” in the traces to be excluded, or in our case whether the network can recognize the non-specific bindings in the mutant groups from true target binding, and assign a low score. To evaluate the sorting task at an ensemble level, the scores of all traces from training are plot into a count histogram (Figure 4.4a). The mutant groups have traces with both high and low scores while the traces in wild type groups have only low scores assigned. It means the network can find the non-specific binding traces or “wild type-look” traces in the mutant groups. The distribution of high-score traces and low-score traces are well separated and only very few traces have scores near the threshold edge 0.5. For that reason, the trained network is smart enough to easily distinguish the “good traces” and “bad traces”, and properly set them apart to “Accept” or “Reject” groups. In practical use, if the researchers want to have higher certainty for accepted traces, they can set the score threshold higher than 0.5 to 0.6, 0.7, or even higher, with the trade-off as reducing the number of accepted traces; if the researchers want to include more accepted traces and/or are more tolerant of false positives, they can set the score threshold lower than 0.5. For the discussion in this thesis, the score threshold is 0.5 and traces to be accepted have scores higher than 0.5.

Rather than simply using the machine learning as a “dark box”, we tested the trained LSTM network on simulated data (Figure 4.4b-d) to reveal what features and information the network is extracting during the training. Signal-to-noise is an essential criterion in sorting traces. The noise level of the T790M experiment data used in training network is around 0.15. Not surprisingly, the trained network accepted more simulated traces that have similar noise level than simulated traces with higher noise level (Figure 4.4b). The network might become sensitive to signal-to-noise when it’s far off the SNR of training dataset and therefore doesn’t accept traces with the high noise level. To examine whether the network can extract the kinetic information contained in the traces, we tested the trained network on simulated data that was generated from a uniform distribution of rate constant. If the network can get kinetic information from training dataset during the training process, the trained network should be able to sort out the simulated traces having similar kinetic information from the whole simulated traces pool. τ_{bound}, $\tau_{unbound}$ and N_{b+d} are the three most important metrics related to the kinetic binding process. When plotting these three parameters obtained from the simulated traces accepted by the trained network into the N_{b+d} vs. τ_{bound} chart and $\tau_{unbound}$ vs. τ_{bound} chart(Figure 4.4c, d), a pattern of the distribution that is similar to the distribution of the experiment data[31] used in training can be observed. All these validations on simulated traces increased the reliability of using machine learning as a new tool in sorting traces from the kinetic fingerprinting measurement.

To compare the performance of machine learning algorithm and HMM thresholding, we tested the trained LSTM network on LOD experiment data from T790M mutation detection[31](Figure 4.5). The LOD experiment conditions are different in the

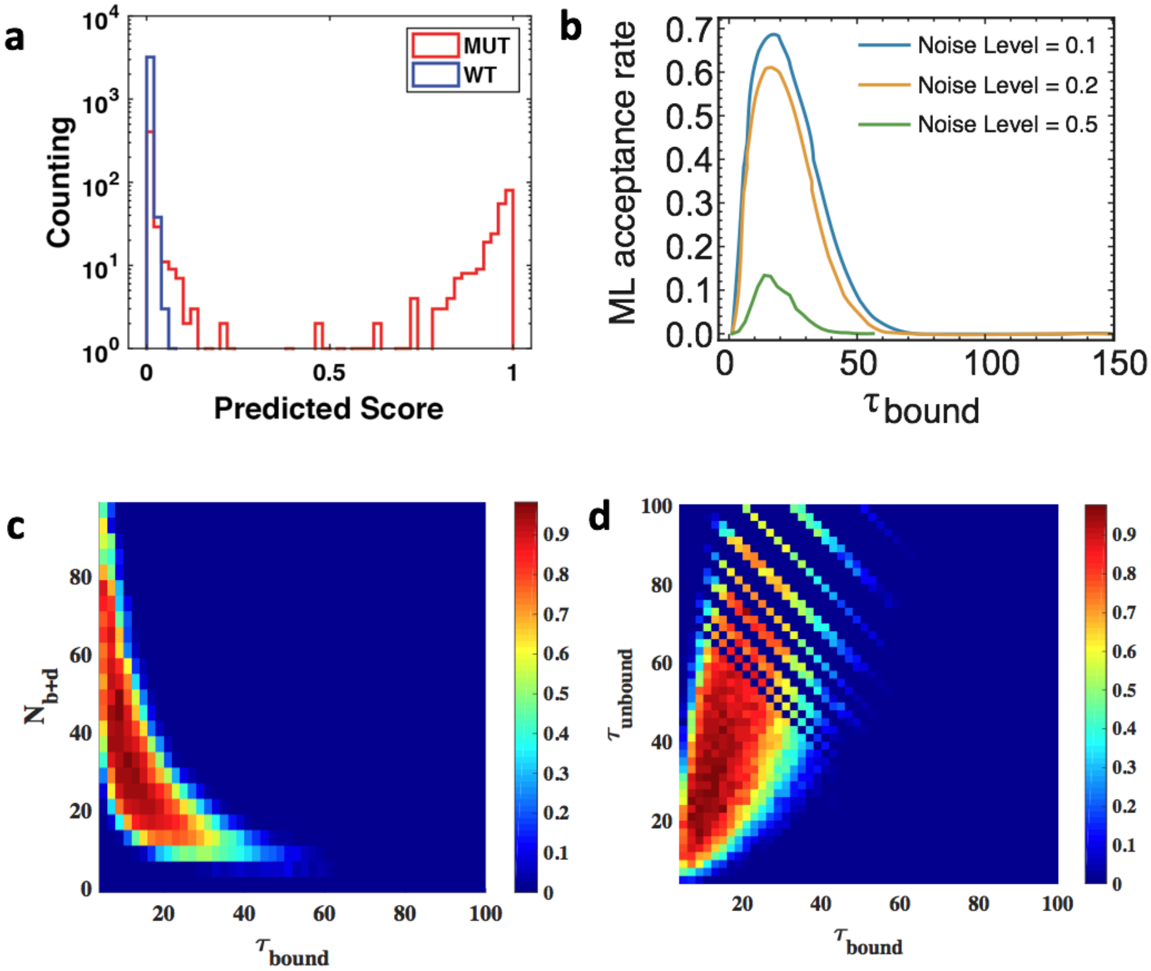


Figure 4. 4. Validation of classification by the machine learning algorithm using simulated traces. **a,** score distribution from machine learning assignment; **b,** *τ_{bound}* distribution of simulated molecule traces having different noise level accepted by machine learning algorithm. Traces having similar noise level (0.1 and 0.2) to the traces used in training (0.15) are accepted more than molecule traces group with higher noise level (0.5); **c, d,** results plotted with different parameters of testing trained network on simulated traces that have a uniform distribution of kinetic binding rate constant.

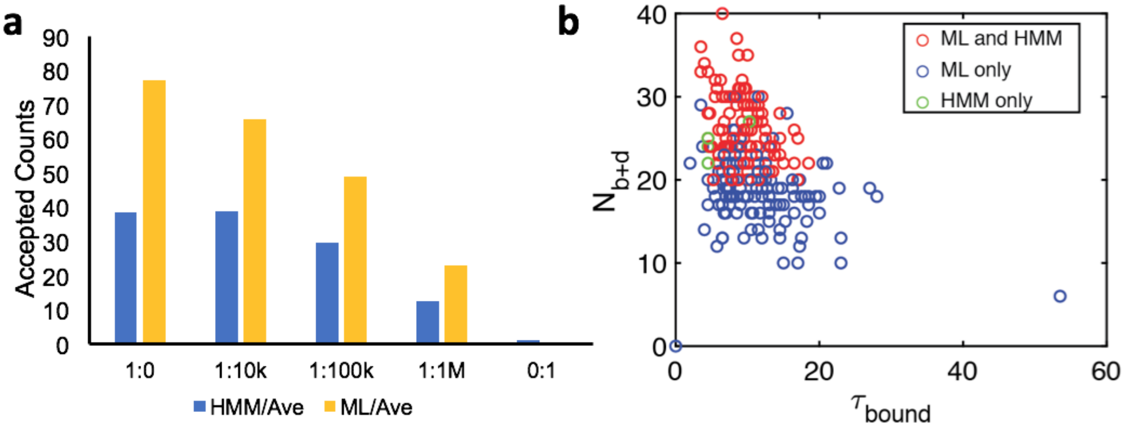


Figure 4. 5. Comparison of HMM thresholding and machine learning classification. **a**, Accepted molecule numbers comparison; **b**, agreement comparison.

concentration ratio of mutant T790M and wild type T790 as 1:10k, 1:100k and 1:1M. Compared to HMM thresholding accepting 38.8, 29.7 and 12.7 traces as the average number in each movie under each ratio condition, the LSTM network accepted 66, 49 and 23 traces respectively (Figure 4.5a). The increasing fold of the number of accepted traces is 1.70, 1.65 and 1.81. Under experiment conditions when no wild type (ratio noted as 1:0) present and no mutant (ratio noted as 0:1) present, the HMM thresholding accepted 38.7 and 1.1 traces as average number in each movie, while LSTM network accepted 77.3 and 0.125 traces respectively. Note that these two conditions experiment data were used in the training. So the number increasing fold of the mutant only groups (ratio noted as 1:0) being higher as 1.99 is not unexpected due to the reason that the network should have better classification performance on data it is trained with. In the wild type only groups, the LSTM network accepted 1 trace across 8 movies. Compared to 9 traces accepted by HMM thresholding from the same dataset, the machine learning algorithm yielded almost zero backgrounds in control groups. The traces number here directly stands for the number of molecules detected by kinetic fingerprinting under the corresponding experimental condition. Across all the experiment conditions in T790M LOD detection, the LSTM network accepted more molecules than HMM thresholding, plus yielding fewer false positive in control groups. Accordingly, using the machine learning algorithm in SiMREPS measurement may increase the sensitivity of detection at low concentration, when only very few molecules can be found on the imaging surface and might not pass the HMM thresholding.

Besides the accepted molecule number comparison, the agreement between HMM thresholding and machine learning algorithm on classification was also investigated.

When plotting the accepted molecule from T790M detection mutant only groups in the N_{b+d} vs. τ_{bound} chart in the style that molecules accepted by different methods in different colors (Figure 4.5b), we found that the machine learning algorithm has a good agreement with HMM thresholding. Almost all molecules accepted by HMM thresholding have also been accepted by the machine learning algorithm (colored in red). Only four out of 116 molecules are accepted by HMM thresholding but not by the machine learning algorithm (colored in green). There are some molecules not accepted by HMM thresholding but accepted by the machine learning algorithm (colored in blue) having overlapping positions with red circles on that plot. We guess that these molecules got screened out during the HMM thresholding, even though they might have similar binding kinetics as the molecules passed the thresholding, as indicated by the similar N_{b+d} and τ_{bound} values on the plot. More investigation of those traces will be needed. Overall, the machine learning algorithm is more tolerant than HMM thresholding in molecule traces sorting, and it can include almost all the selected molecule traces by HMM thresholding. We also tested the trained network on the experiment data from T790M standard curve measurement[31]. The coefficient of determination(R^2) fit on traces number from LSTM selection is 0.9947, while the HMM thresholding gives 0.9902.

Although further detailed check of the disagreed traces might be needed, the current results of using machine learning algorithm in analyzing single channel data from SiMREPS measurement show that LSTM network sorting has the potential to outperform the current HMM thresholding, thus assisting SiMREPS measurement to increase detection sensitivity.

4.3.2 Use machine learning to analyze double channel data from single molecule FRET measurement

As mentioned earlier, the current analysis of smFRET data involves manual selection of traces. This manual selection is time-consuming and introducing inconsistency over time. In most cases, researchers only include the dynamic region of the traces for further analysis, for example, the segment with cross-correlation between donor and acceptor signals. Therefore, the task for double channel data analysis has an extra step of segmentation in addition to the sorting. Here, we also used LSTM neural network for classification similar as in single channel data analysis. Since LSTM neural network is sensitive to information change from time point to time point, it can do segmentation as well. Figure 4.7 shows representative segmentation from LSTM.

In most cases, the region selected by the network (box colored in light blue) has good agreement with the region selected by a human (box colored in dark blue). In some cases, disagreement happens because the network is susceptible to sudden signal changes such as fluorescence blinking, in which case the network might have taken as a photobleaching event. Humans are much “smarter” in those cases to skip the blinking part. Together with the sorting task, the highest concordance between human selection and machine learning algorithm achieved by LSTM network is 89.8%. Since there are no ground truth labels from experiments as in the single channel data scenario, we use the concordance to evaluate the performance of the network compared to human beings. The concordance is calculated as the ratio of the number of traces agreed on both human and algorithms (pick and drop) divided by the number of all traces included.

To increase the agreement between human selection and algorithm, or in other words to train the network to be more like a human, more information to be collected during training is needed. One very crucial criterion during human selection is the degree of anti-correlation between donor and acceptor signals. We creatively converted the 1-dimensional time series data into 2-dimensional intensity plot (Figure 4.8). Each trace was converted into an image with donor intensity and acceptor intensity as the axes. By doing this, a frame-to-frame comparison task of the time series data is converted into an image recognition task. As shown in the figure, the cross-correlation degree can be reported as the pattern of the spots on the images. In the diagonal direction, a slope of -1 pattern will show up if there is cross-correlation region in the traces. We used AlexNet, a well-known CNN network for the image classification task. When connecting the LSTM and AlexNet together into one big network as a traces selector, the highest concordance between human selection and network sorting is 96.1%. The training and test included 9 datasets from three users, while training and test were on different a dataset. From the results, the use of CNN does increase the performance of the traces selector. The drawback of using CNN is the need of computing power cannot be satisfied with personal laptop or desktop. All the CNN trainings in this work were done on GPU at Amazon Cloud. In our user interface of the automatic selector being currently designed, the CNN training is disabled in default. Users may use the Walter lab data pre-trained AlexNet for classification purpose. The workflow of automatic traces selector is shown in Figure 4.8. The LSTM network is capable for both traces segmentation and traces classification. CNN network can be coupled in when computing power permits to boost the classification concordance.

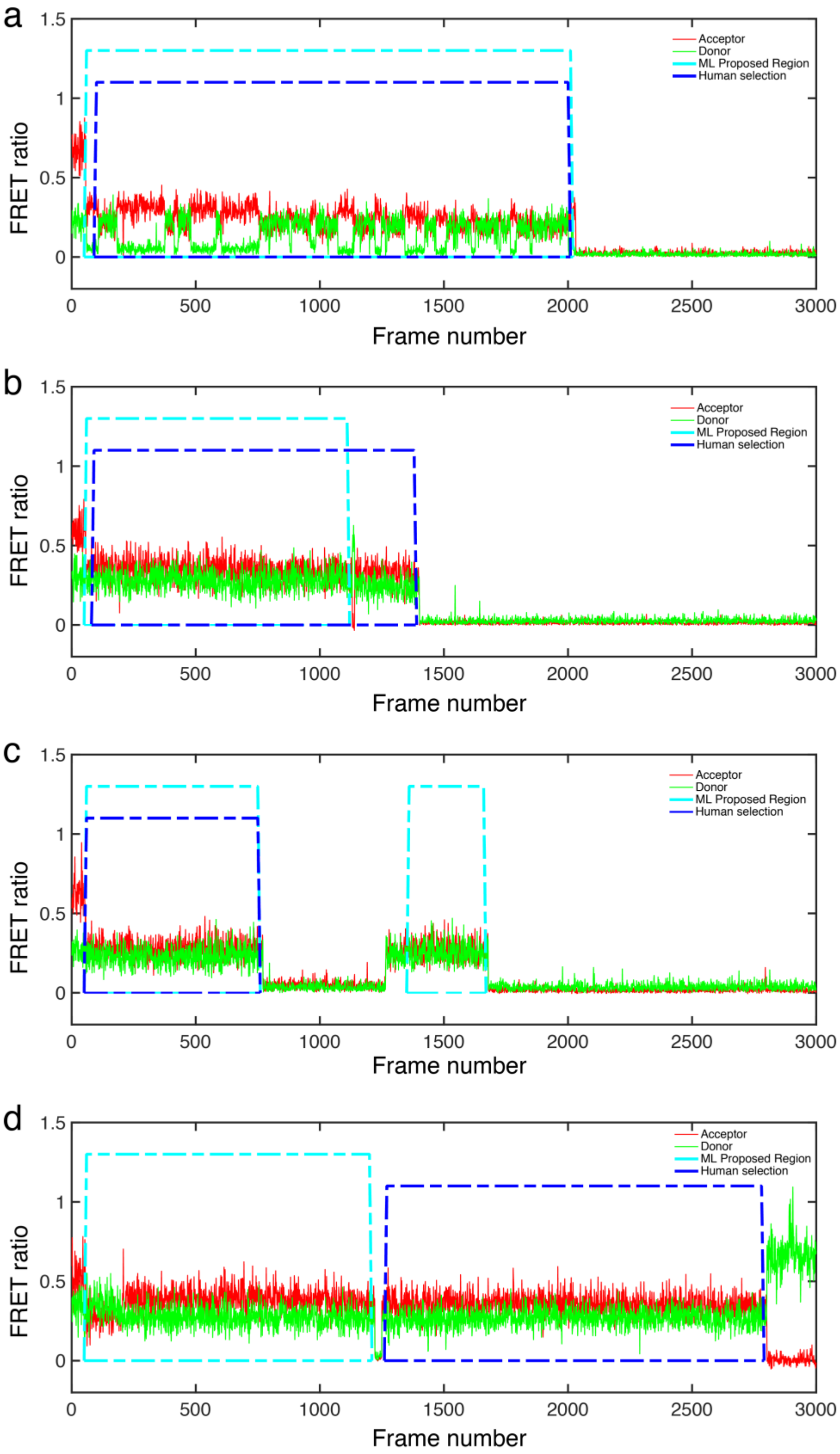

a
b
c
d
FRET ratio
Frame number
0
0.5
1
1.5
0
500
1000
1500
2000
2500
3000
Acceptor
Donor
ML Proposed Region
Human selection

Figure 4. 6. Representative examples of segmentation comparison between human selection (dark blue box) and machine learning algorithm (light blue box). **a**, the machine learning algorithm selects almost the same region as human selection; **b**, human selection covers more region than machine learning algorithm because the algorithm is sensitive to sharp signal change and sometimes may take it as an intensity blinking; **c**, the machine learning algorithm can recognize and skip the photo-bleach-like region, and selects more regions for further analysis than human selection; **d**, rare case does happen as algorithm selection and human decision are nothing in common. Reasons can be that the algorithm stopped before the sharp signal drop while human chooses to skip that signal-drop region.

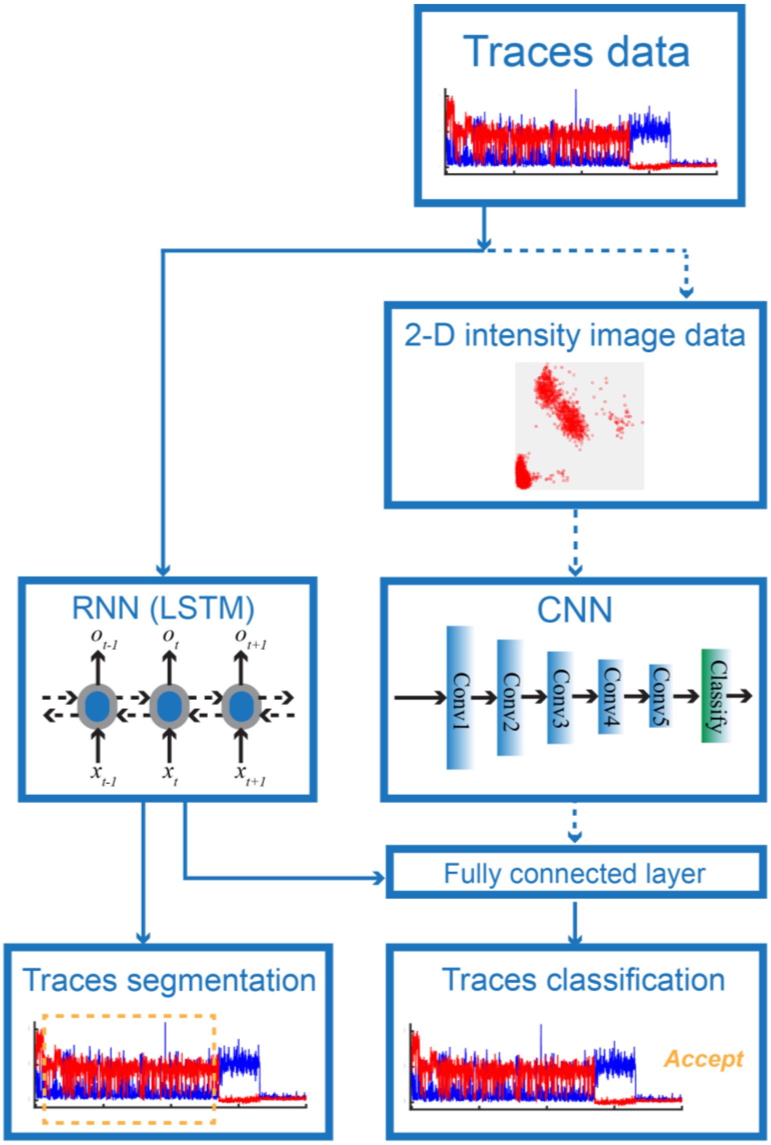


Figure 4. 7. Workflow of using machine learning to analyze double channel data. For laptop users who don't have GPU installed, the dashed line pathway will be disabled in the user interface. LSTM is the main network used in both sorting and segmentation.

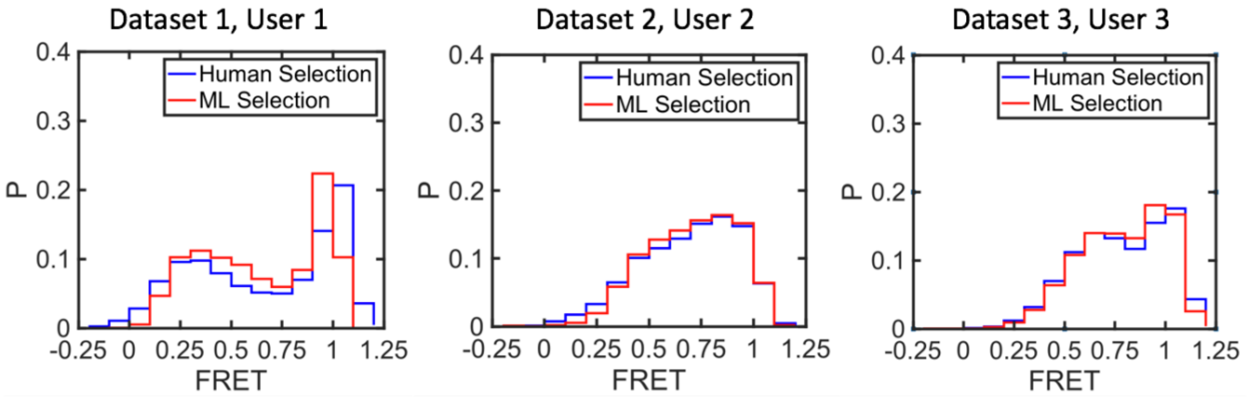


Figure 4. 8. Comparison between human selection and machine learning algorithm of all the traces. Each user used different dataset therefore having different manual selection criteria.

To evaluate at an ensemble level of how different human selection and algorithm sorting are, we plot the histogram of FRET value from 3 users' 3 datasets in Figure 4.9. This type of histogram is very often used in daily smFRET data analysis after manual selection to get the FRET value distribution in the molecule population. The results suggest that for some dataset like #2 and #3 in Figure 4.9, the traces selection result from machine learning algorithm is indistinguishable from the human selection. While some dataset might involve more heterogeneous information like #1, disagreement between human and machine learning algorithm becomes more evident. We believe that increasing the training data size and data diversity would diminish this issue. Overall, the machine learning algorithm can sort the double channel traces from smFRET measurement in ~90% similarity to human selection, without the need of spending hundreds of hours. We also believe the optimization of network training might increase the concordance to better performance.

4.4 Discussion

In this chapter, I introduced the idea of using machine learning algorithms to assist data analysis in single molecule measurement. Long short-term memory (LSTM) neural network was used for its unique capability of analyzing time series data. In single channel data analysis, LSTM network can achieve a higher selectivity sensitivity than traditional HMM thresholding. The validation conducted by testing on simulated traces suggests that the LSTM network can get kinetic features from single molecule traces during training and recognize them during classification. In double channel data analysis, the LSTM network can sort traces as well as segment out the region of interest

(ROI). Coupling the AnexNet into LSTM network can increase the performance of the automated selector, although it is limited to the practical computing power. In addition to getting similar final results as human selection, the automated selector can finish the task in several minutes, saving endless hours on manual selection for researchers. We suppose the productivity and efficiency of data analysis in smFRET measurement can be massively increased with this innovative automated selector.

For better performance of the network in sorting, we believe increasing the training data size to include more diverse data will be a simple but valuable way. Increasing the layer number of the network will theoretically increase the network's performance. However, the drawback is higher demanding of computing power or time for training. Other machine learning or deep learning algorithms may also be good choices.

We hope this work can contribute the data analysis tool kit in single-molecule measurement and introduce machine learning as a new option to solve research problems in biophysics field.

Chapter 5

Summary and Outlook

5.1 Summary of results

Over the past decades, the development of single molecule fluorescence microscopy has enabled researchers to observe the microscopic world at unprecedentedly high spatial resolution. Measurement of a single molecule removes the ensemble average from traditional bulk observation, allowing the exploration of concealed heterogeneities in complex systems and the direct detection of dynamic state changes without synchronization. No matter what variables are measured, the time dependence of the parameter can generate information about many dynamic processes, such as the duration of excited states, local environmental fluctuations, enzymatic activity differences, etc. This unique feature gives an entry point to engineer dynamic behaviors of molecules.

The DNA nanotechnology field is well-known for its focus and ability to assemble nucleic acids into nano-devices that have specific engineering functions such as building blocks, cargo sorting and energy transport[29,120–122]. The power to provide information of heterogeneous behaviors of molecules makes single molecule microscopy an ideal tool for optimization of nano-device design. Single molecule techniques provide an invaluable method for observing the behavior of molecules individually. Apart from using single molecule microscopy to characterize self-assembled DNA structures to optimize their design[123], the temporal information recorded over the observation provides insightful knowledge about dynamic behaviors of DNA nano-devices. One example is using smFRET to characterize how DNA walkers translocate in the microscopic world as I described in Chapter 2.

In Chapter 2, a new type of movement mechanism of DNA walkers was introduced. Our novel DNA "acrobat" is moving in cartwheeling fashion, achieved through toehold exchange displacement between foothold strands. The movement is completely autonomous, contrasting with previously reported DNA motors that require energy supply, which provides a new option for potential unbiased molecular transport. smFRET measurements were used to provide stepping information of various DNA walker designs. Through the interpretation of lifetime changes in different dynamic states, the rate-limiting step during the translocation was identified as the toehold association and dissociation. This mechanistic characterization in return helped modify the design of DNA walkers to obtain the fastest one. The fastest toehold exchange DNA walker is termed $\boldsymbol{W_{6\_13\_6}}$, which has a stepping rate constant approaching 1 s^{-1}, 10- to 100-fold faster than prior DNA walkers. Through super-resolution imaging and single particle tracking using TIRF microscopy, we demonstrated that this new DNA walker moves via the same mechanism on two-dimensional substrate surfaces. The DNA walker moves over hundreds of nanometers within 10 minutes, in quantitative agreement with the prediction from stepping kinetics measured by smFRET.

Besides providing information on heterogeneities in the microscopic world, the high spatial resolution single molecule measurements can achieve make them a powerful tool to find a specific type of single molecule in a pool of molecules, just like searching for a needle in the haystack. This idea was demonstrated by the kinetic fingerprinting technique- SiMREPS, a robust detection technique capable of extremely low concentration detection through recording the transient binding of fluorescence probes to target molecules. In Chapter 3, I systematically discussed matters needing attention during SiMREPS experiments, including fluorescent probe design, sample preparation, imaging conditions and data analysis. I also discussed how different experimental set-ups might impact the measurement. In addition to improvement of the method itself, expanding the detectable targets spectrum also is of interest. Currently, the targets include miRNA and ctDNA[30,31]. Exploration of using SiMREPS measurement to detect epigenetic mutations such as CpG site methylation was explored.

In all single molecule experiments, the data analysis process is always of great importance to extract the enormous information collected during observation. The manual selection of traces after smFRET measurement is usually time consuming. To improve the productivity of research efforts, machine learning algorithms were introduced in Chapter 4. Our RNN and CNN coupled neural network yields a 96% concordance with human selection, without the need for hundreds of hours devoted to manual selection. In a similar application of machine learning to analyze single-channel data from SiMREPS measurements, the network outperformed current HMM thresholding in a test on T790M LOD experimental data. This improvement of data analysis may help SiMREPS measurements achieve a higher detection sensitivity. In general, the application of machine learning algorithm to single molecule data analysis introduces new options and opportunities for researchers.

Taken together, the above work shows how single molecule fluorescence microscopy is powerful and robust in both characterizing DNA nano-devices and diagnostic detection of nucleic acids. Common between these applications is the dynamic information collected through single molecule measurement. To better analyze the dynamic information, machine learning algorithms were introduced to the single molecule field.

5.2 Outlook

The new type of autonomous DNA walker presented in Chapter 2 suggests that substantial improvements in the operating rates of broad classes of DNA nanomachines utilizing strand displacement are possible. Further optimizing the mechanical properties of the system may also be possible, such as the reach of the walker, entropic tension in single-stranded linker segments, etc. To surpass the current “speed limit”, dynamic DNA nanotechnology may need to incorporate further innovations such as more precise control of local DNA mechanics and energy transduction. The ultimate goal of developing DNA walkers is to mimic natural biomotor transportation phenomenon in more controllable fashion. Besides molecular transport, other tasks such as sorting and target searching are also promising directions. Random walk as a classical model is widely used in computer science algorithms for searching tasks. The toehold exchange DNA walker developed here utilizes a random walk mechanism. An attempt to achieve a goal searching task using this DNA walker was made in my research. In the single particle tracking on 2D surface experiment, we attached a Cy5 fluorophore dye on some foothold strands and a Cy3 fluorophore dye on the DNA walker strand. When the DNA walker finds the Cy5 labeled foothold, a FRET signal should be measurable due to energy transfer. However, the observation of success in this goal search task yielded inconclusive results because of limits in the experimental conditions: the leakage of Cy3 signal obscured the FRET signal. If some optimization of the experiments can be achieved, the searching capability of this random DNA walker should be demonstrable. Once the goal searching and fluorescence resonance energy transfer can be realized, molecular transport using this DNA walker can be attempted[124]. For example, instead of labeling with a fluorophore dye, molecules that need to be transported can be attached to the DNA walker, such as coenzyme cargos. Similar autonomous DNA walkers using a different strand displacement mechanism may also be considered. If needed, coupling energy to speed up the toehold exchange DNA walker can also be explored, such as introducing polymerase reaction.[125] Although the translocation system may become more complex and as a result harder to control, coupling energy will earn the potential velocity increment and greater controllability in moving directions. Another interesting direction to explore is changing the experimental temperature. All the smFRET measurement and single particle tracking in chapter 2 were conducted under room temperature. The locomotion speed of the DNA walkers correlated to the kinetic binding information may vary at different experimental temperatures[126–128]. Higher experimental temperature may lead to faster stepping rate since the rate-limiting step during toehold exchange displacement is the strands binding.

The discussion in Chapter 3 suggests a further optimization of the SiMREPS method itself can be achieved from multiple directions. Increase of mass transfer of targets in solution to the imaging surface would push the detection concentration limit of

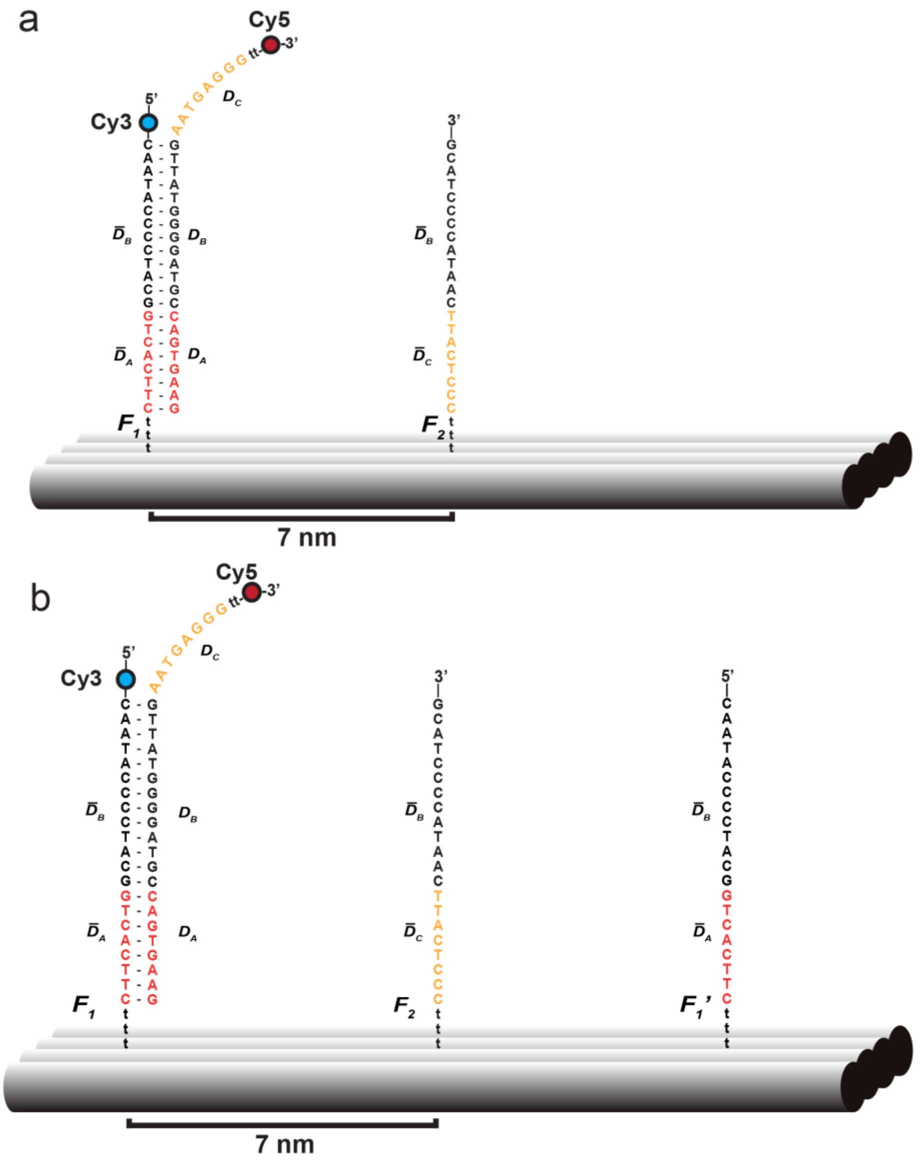
SiMREPS to a lower level. An adjustment of the imaging buffer to include formamide would change the binding kinetics of nucleic acids. Formamide presence should decrease the hybridization stability of fluorescent probes and targets, thus shortening the binding cycle time. This adjustment would make the detection of longer sequences by SiMREPS possible. The current length of the fluorescence probe is 8-9 nucleotides. A greater length of the binding region on the target would increase the selectivity in detecting genomic mutations. Furthermore, expanding the binding pairs from nucleic acid-nucleic acid to other molecules will massively broaden the analyte spectrum, such as protein-protein binding, where the fluorescence probe and target will be, for example, both proteins. The CpG methylation detection discussed in Chapter 3 offers the option of using nucleic acid-protein binding pairs in SiMREPS, thus pushing the detection limit from single nucleotide to single functional group differences. Due to the significant role CpG methylation plays in epigenetic gene regulation[129,130], various detection methods have been developed[131]. Most of these methods focus on the fraction of methylation in the gene sample at a bulk level. The precise position of the CpG methylation in the sequence is of great importance to understand the gene regulation process. This type of high sequence specificity can be achieved during SiMREPS detection through careful sequence design of capture probe and fluorescence probe, making SiMREPS a promising detection tool for CpG methylation.

The use of machine learning in single-molecule data analysis discussed in Chapter 4 was considered as a new approach to reduce the human subjectivity introduced during the analysis. More and diverse training datasets should be included to build the network to be more “intelligent” and even less biased. The usual concern of using machine learning algorithm in practice is the abstract and sometimes unknown features that the network is learning. More investigation into the features detected by our algorithm will be beneficial to the whole single molecule science community. Ideally, time series data from any other experimental measurements that contains special pattern of transitions should be handleable by our machine learning algorithm. Application of machine learning to other types of single molecule datasets is therefore very promising.

The power of a technique becomes unlimited when people think creatively about how to use it. Determine the problem and invent the approaches to solve it is what scientists have been doing. My journey has just begun.

APPENDIX

Supplementary Information to Chapter 2



c

$W_{8\_13\_8}$ 5'-GAAGTGACCGTAGGGGTATTGAATGAGGGtt-Cy5-3'

$W_{7\_13\_7}$ 5'-AAGTGACCGTAGGGGTATTGAATGAGGtt-Cy5-3'

$W_{6\_13\_6}$ 5'-AGTGACCGTAGGGGTATTGAATGAGtt-Cy5-3'

$W_{5\_13\_5}$ 5'-GTGACCGTAGGGGTATTGAATGAtt-Cy5-3'

Figure A1.1. Design of cartwheeling walker and DNA tile**. a,** ***2-Foothold*** system foothold strands **F_1** and **F_2** are 5' and 3' extensions of ssDNA strands within the 4-helix tile (grey cylinders). **b,** ***3-Foothold*** system with **F_1'** having same sequence as **F_1**. **c,** The walker ***W*** is a single-stranded DNA oligonucleotide comprising a 13-nucleotide branch migration domain ***D_B*** (coloured black) flanked by two 5- to 8-nucleotide toehold domains ***D_A*** (coloured red) and ***D_C*** (coloured orange) with distinct sequences.

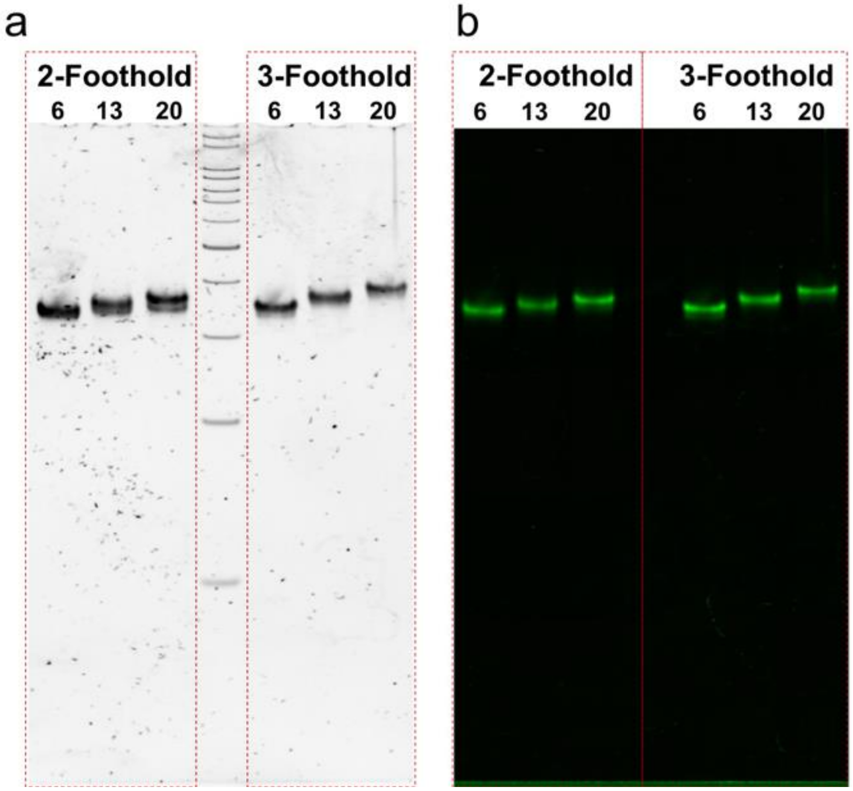
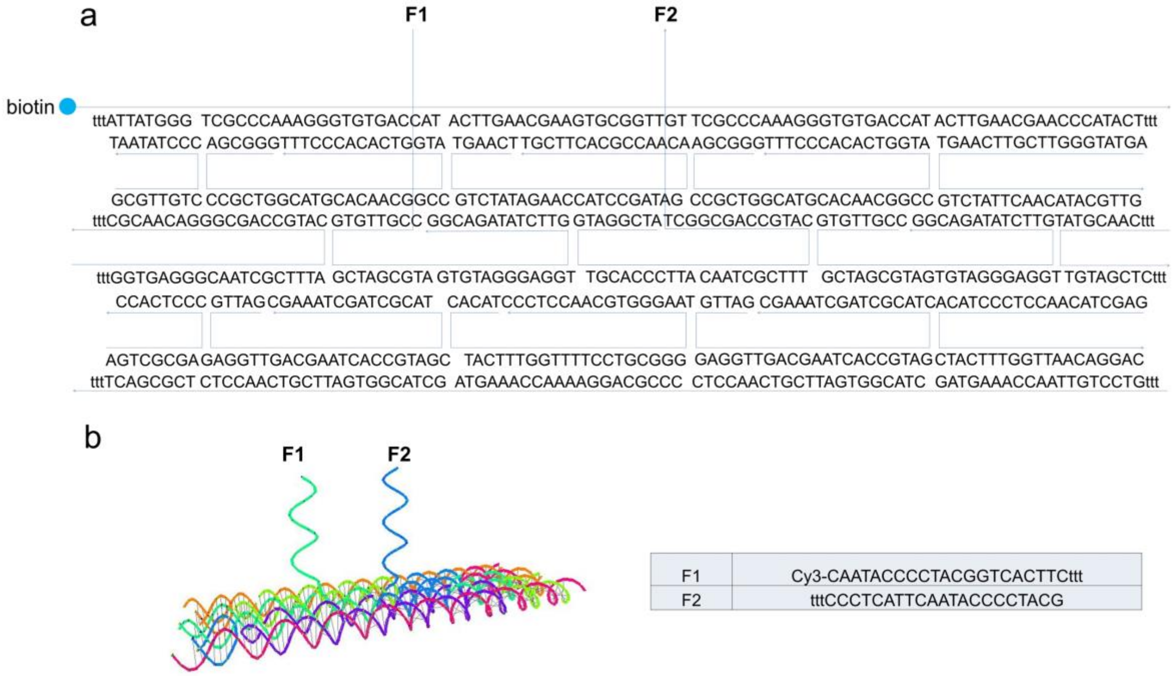


Figure A1. 2. 5% Native PAGE characterization of *2-Foothold* and *3-Foothold* DNA tile systems**. a,** 5% native PAGE with SYBR Green stain of different tile constructs used in the paper. **b,** Fluorescence gel characterization of Cy3-labeled tile. The number above each lane (6, 13, 20) represents the number of nucleotides in the middle domain ($\overline{\boldsymbol{D_B}}$) of each foothold strand.


a
F1
F2
biotin
tttATTATGGG TCGCCCAAAGGGTGTGACCAT ACTTGAACGAAGTGCGGTTGT TCGCCCAAAGGGTGTGACCAT ACTTGAACGAACCCATACTttt
TAATATCCC AGCGGGTTTCCCACACTGGTA TGAACT TGCTTCACGCCAACA AGCGGGTTTCCCACACTGGTA TGAACTTGCTTGGGTATGA
GCGTTGTC CCGCTGGCATGCACAACGGCC GTCTATAGAACCATCCGATAG CCGCTGGCATGCACAACGGCC GTCTATTCAACATACGTTG
tttCGCAACAGGGCGACCGTAC GTGTTGCC GGCAGATATCTTG GTAGGCTA TCGGCGACCGTAC GTGTTGCC GGCAGATATCTTGTATGCAACttt
tttGGTGAGGGCAATCGCTTTA GCTAGCGTA GTGTAGGGAGGT TGCACCCTTA CAATCGCTTT GCTAGCGTAGTGTAGGGAGGT TGTAGCTCttt
CCACTCCC GTTAG CGAAATCGATCGCAT CACATCCCTCCAACGTGGGAAT GTTAG CGAAATCGATCGCATCACATCCCTCCAACATCGAG
AGTCGCGA GAGGTTGACGAATCACCGTAGC TACTTTGGTTTTCCTGCGGG GAGGTTGACGAATCACCGTAG CTACTTTGGTTAACAGGAC
tttTCAGCGCT CTCCAACTGCTTAGTGGCATCG ATGAAACCAAAAGGACGCCC CTCCAACTGCTTAGTGGCATC GATGAAACCAATTGTCCTGttt
b
F1
F2

F1	Cy3-CAATACCCCTACGGTCACTTCttt
F2	tttCCCTCATTCAATACCCCTACG

Figure A1. 3. DNA sequence design for 4HX tile with *2-Foothold***. a,** The structure incorporates two ssDNAs as the two footholds, F1 (5’-CAATACCCCTACGGTCACTTC) and F2 (CCCTCATTCAATACCCCTACG-3’). The distance between 2 footholds are designed to be 7 nm and facing the same side of 4HX tile. **b,** Computer modelling (Tiamat) of DNA nanostructure and the detailed sequence and labelling strategy of T1 and T2. Cy3 dye is labelled at 5’ of F1 with 2 T bases as spacer. For both F1 and F2, A single-stranded 3T spacer was added between the foothold and the tile to allow for flexibility.

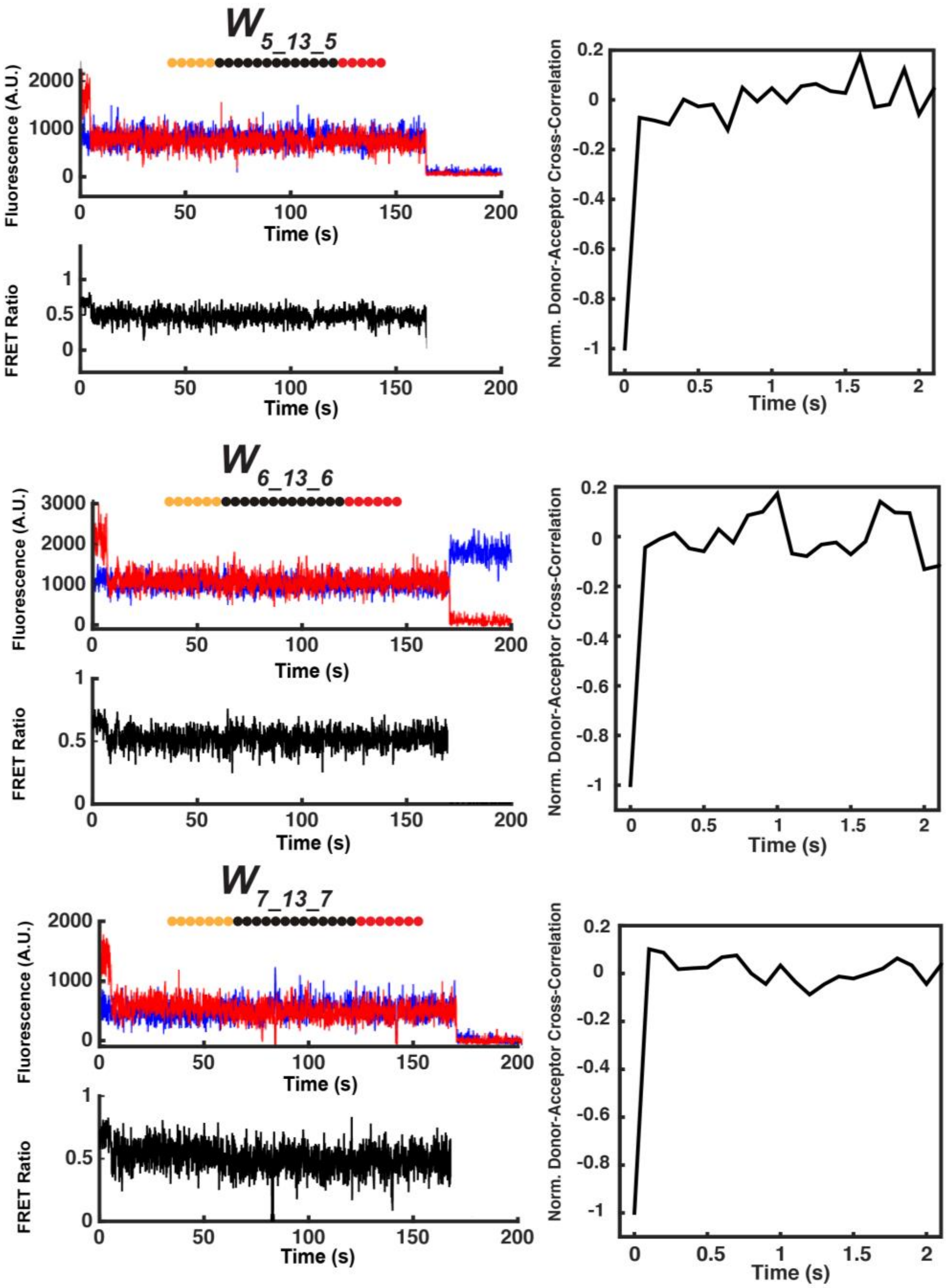


Figure A1. 4. Evidence of rapid FRET dynamics for $W_{5\_13\_5}$, $W_{6\_13\_6}$, $W_{7\_13\_7}$ on *2-Foothold* DNA tile**.** Rapid anti-correlated fluctuations in Cy3 (blue) and Cy5 (red) fluorescence intensity for a single walker-tile complex, suggestive of branch migration in hybrid state $\boldsymbol{S_{1+2}}$.

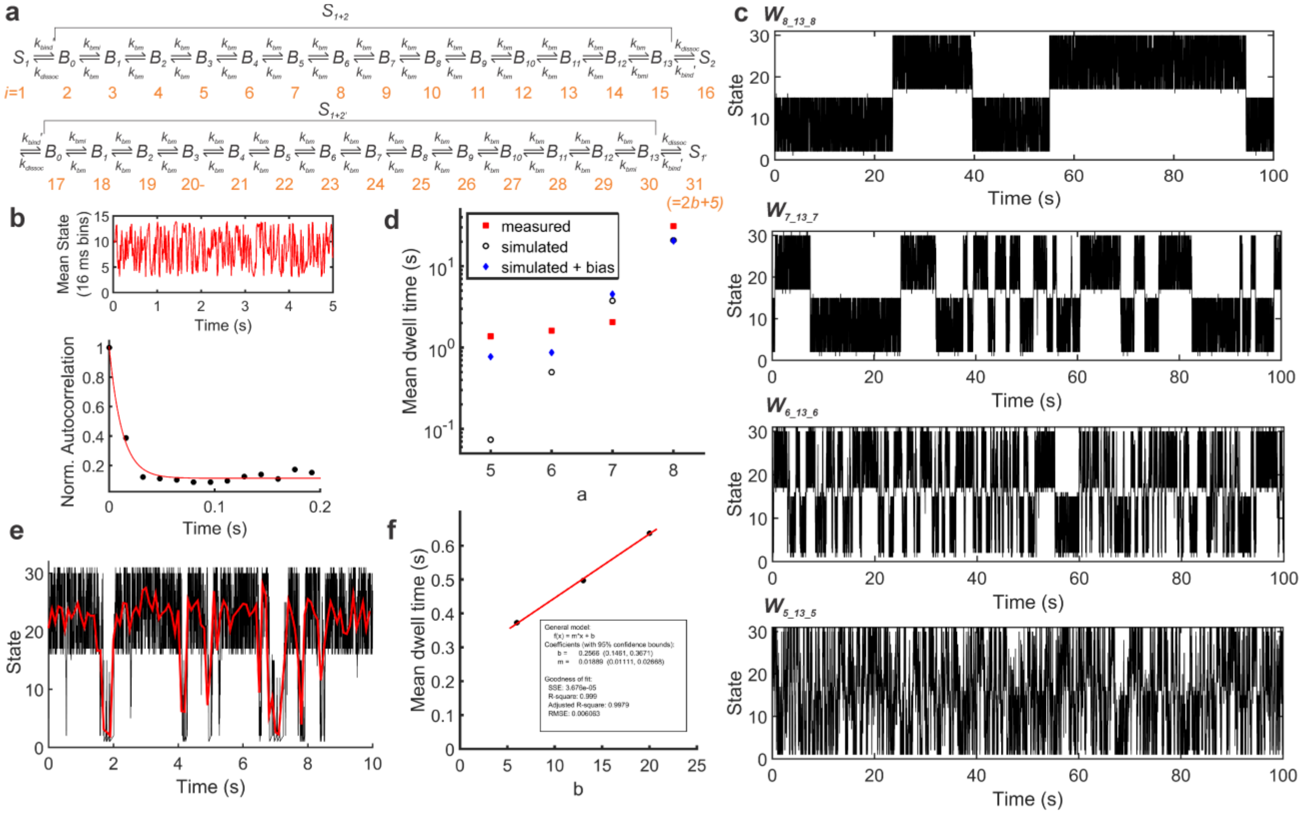
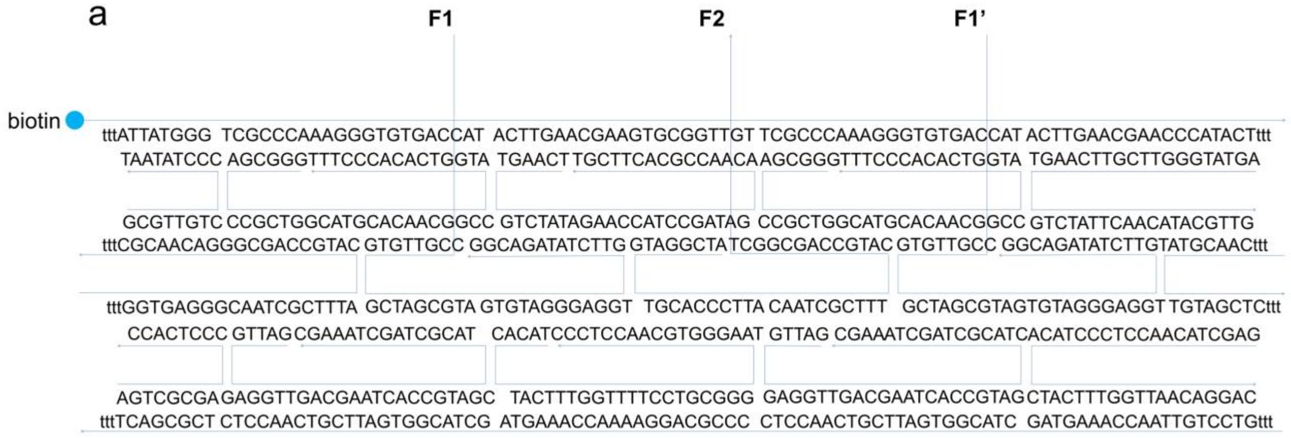
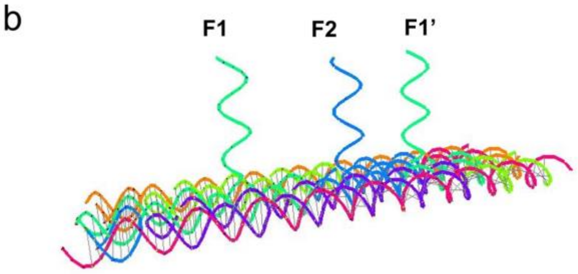


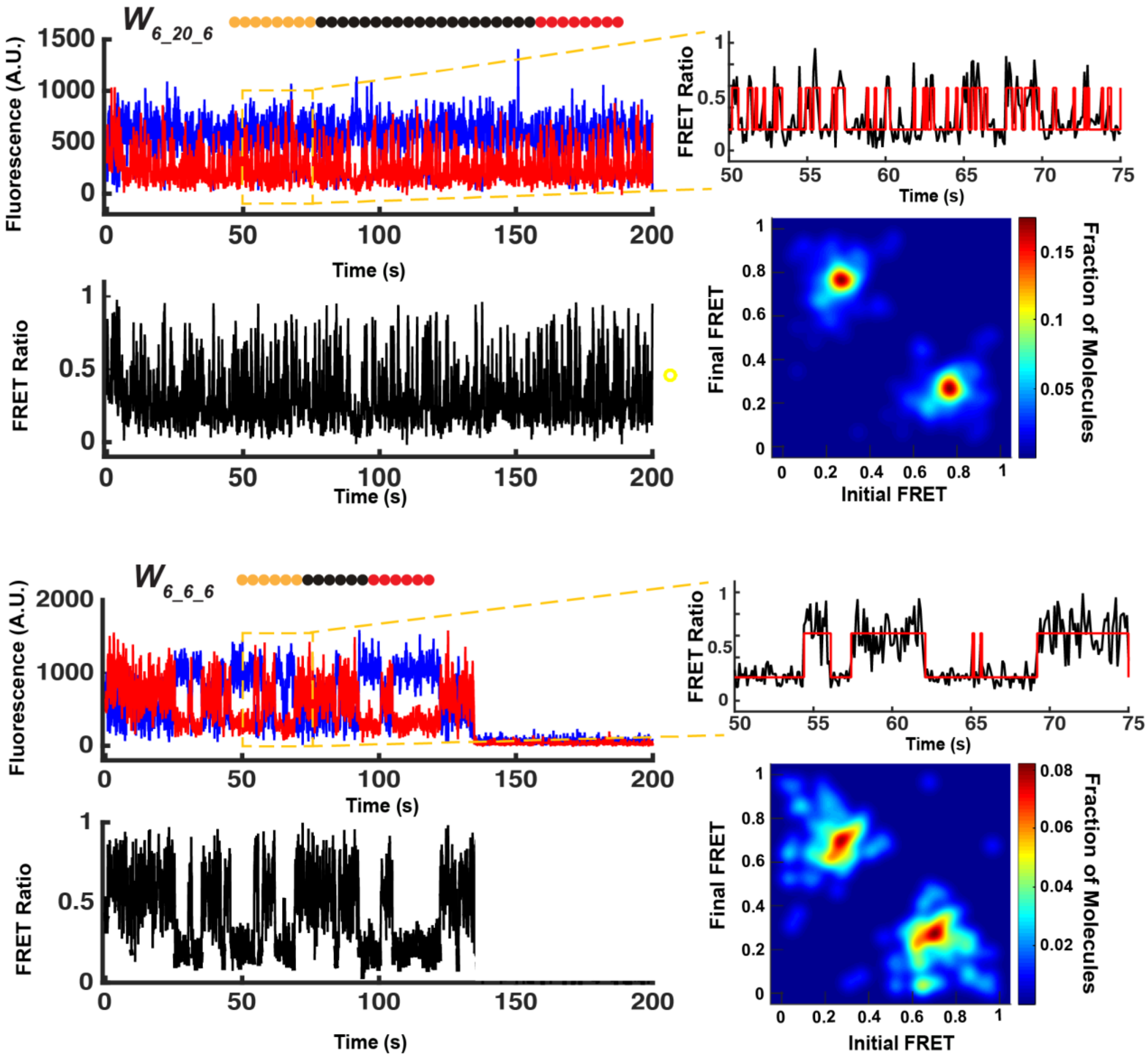
Figure A1. 5. Monte Carlo simulation of cartwheeling DNA walkers in a 3-foothold system**. a,** Scheme for kinetic modelling of 3-foothold system (***b*** = 13 nucleotides in the depicted scheme). See Supplementary Note 1 for details regarding the model. **b,** (top) Representative portion of a simulated trajectory of $W_{8\_13\_8}$ in a 3-foothold system, zoomed in to show the rapid fluctuations among branch migration states. State values in this plot are binned to a time resolution of 16 ms to match the time resolution of donor-acceptor anticorrelation measurements (see Fig. 1, main text). (bottom) Exponential fit to the normalized autocorrelation function of the time-binned trajectory shown at the top. The lifetime of the exponential fit is 12.1 ms (95% confidence interval: [9.6, 14.6 ms]). **c,** Representative state vs. time trajectories for simulated walkers with ***b*** = 13 and toehold length ***a*** varying from 5 to 8 nucleotides. Rapid fluctuations among branch migration intermediates are punctuated by rare toehold dissociation and stepping events, which become more frequent as ***a*** decreases. **d,** Mean stepping dwell time of simulated trajectories (black filled circles) with ***b*** = 13 and varying ***a***, as compared to the experimentally determined values (red squares) and simulated trajectories incorporating a toehold length-dependent bias towards one FRET state (blue diamonds). **e,** Simulated trajectory of $W_{5\_13\_5}$ with a 10-fold bias towards binding one foothold (black), along with a time-binned version of the same trajectory (red). The time binning in the red trajectory is 100 ms, to match the time resolution of smFRET measurements. **f,** Mean stepping dwell time of simulated trajectories with varying ***b*** and constant ***a*** (=6). The trend is well fit by a linear function ($R^2 > 0.99$).


a
F1
F2
F1’
biotin
tttATTATGGG TCGCCCAAAGGGTGTGACCAT ACTTGAACGAAGTGCGGTTGT TCGCCCAAAGGGTGTGACCAT ACTTGAACGAACCCATACTttt
TAATATCCC AGCGGGTTTCCCACACTGGTA TGAACT TGCTTCACGCCAACA AGCGGGTTTCCCACACTGGTA TGAACTTGCTTGGGTATGA
GCGTTGTC CCGCTGGCATGCACAACGGCC GTCTATAGAACCATCCGATAG CCGCTGGCATGCACAACGGCC GTCTATTCAACATACGTTG
tttCGCAACAGGGCGACCGTAC GTGTTGCC GGCAGATATCTTG GTAGGCTA TCGGCGACCGTAC GTGTTGCC GGCAGATATCTTGTATGCAACttt
tttGGTGAGGGCAATCGCTTTA GCTAGCGTA GTGTAGGGAGGT TGCACCCTTA CAATCGCTTT GCTAGCGTAGTGTAGGGAGGT TGTAGCTCttt
CCACTCCC GTTAG CGAAATCGATCGCAT CACATCCCTCCAACGTGGGAAT GTTAG CGAAATCGATCGCATCACATCCCTCCAACATCGAG
AGTCGCGA GAGGTTGACGAATCACCGTAGC TACTTTGGTTTTCCTGCGGG GAGGTTGACGAATCACCGTAG CTACTTTGGTTAACAGGAC
tttTCAGCGCT CTCCAACTGCTTAGTGGCATCG ATGAAACCAAAAGGACGCCC CTCCAACTGCTTAGTGGCATC GATGAAACCAATTGTCCTGttt



F1	Cy3-CAATACCCCTACGGTCACTTCttt
F2	tttCCCTCATTCAATACCCCTACG
F1'	CAATACCCCTACGGTCACTTCttt

Figure A1. 6. DNA sequence design for 4HX tile with *3-Foothold*. **a,** The structure incorporates 3 ssDNAs as the three footholds, F1, F1’ (5’-CAATACCCCTACGGTCACTTC) and F2 (CCCTCATTCAATACCCCTACG-3’). The distance between each two footholds are designed to be 7 m and facing the same side of 4HX tile. **b,** Computer modelling (Tiamat) of DNA nanostructure and the designed sequence of footholds F1, F1’ and F2. Cy3 dye is labelled at 5’ of F1 with 2 T bases as spacer. For all three footholds, a single-stranded 3T spacer was added between the foothold and the tile to induce flexibility.



$W_{6\_20\_6}$ 5'-AGTGACTCCGTATCCATGACGTGAGAAATGAGtt-Cy5-3'

$W_{6\_6\_6}$ 5'-AGTGACTGATCTGAATGAGtt-Cy5-3'

Figure A1. 7. Single-molecule FRET characterization of *$W_{6\_20\_6}$* and *$W_{6\_6\_6}$* on *3-Foothold* DNA tile. Representative smFRET trajectories of ***$W_{6\_20\_6}$*** and ***$W_{6\_6\_6}$*** on ***3-Foothold*** DNA tile are shown with Cy3 fluorescence in blue and Cy5 fluorescence in red. Zoomed-in trajectories (upper-right corner of each panel) show FRET transitions for 25-s segments in greater detail. Transition occupancy density plots (TODPs, lower-right corner of each panel) show the most frequently observed FRET transitions across all molecules.

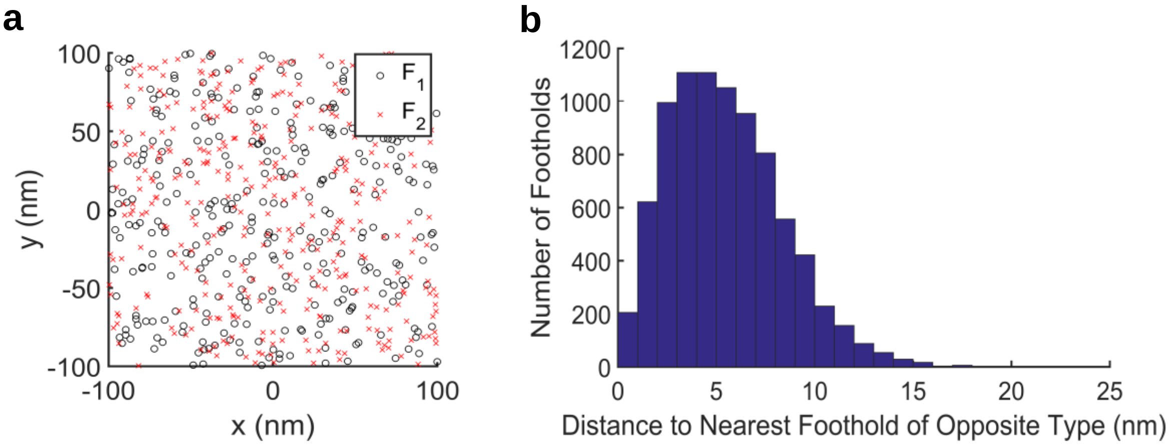


Figure A1. 8. Simulated distributions and distances to nearest neighbour footholds on 2D surfaces. **a,** Representative 200 nm × 200 nm region showing randomly distributed footholds F_1 and F_2. **b,** Histogram of predicted distances to nearest-neighbour footholds of the opposite type within a (1000 nm)^2 region containing 8350 randomly distributed copies each of F_1 and F_2. Foothold positions are assumed to be independent of all other footholds.

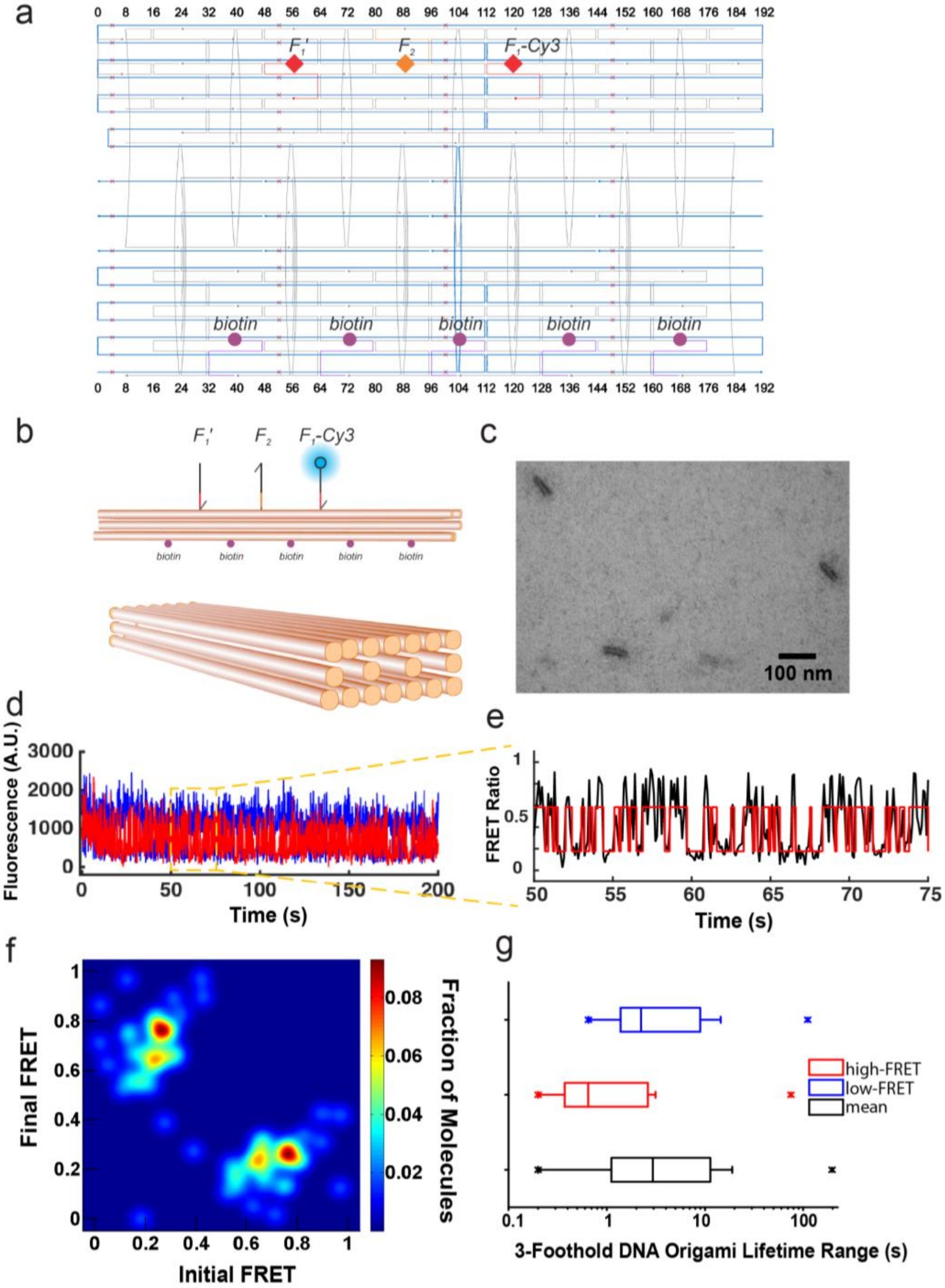


Figure A1. 9. Single-molecule FRET characterization of *$W_{6\_13\_6}$* on *3-Foothold* DNA origami**. a,** caDNAno scaffold routing diagram for ***3-Foothold*** DNA origami, showing positions of footholds and biotins used for anchoring to the imaging surface for TIRF. **b,** Cartoon schematics of ***3-Foothold*** DNA origami, including a side view of foothold and biotin positions (top) and a perspective view of the underlying nanostructure (bottom). The distance between adjacent footholds is predicted to be ~10.5 nm. **c,** TEM characterization of ***3-Foothold*** DNA origami. **d,** A representative single-molecule FRET trajectory of ***$W_{6\_13\_6}$*** on ***3-Foothold*** DNA origami. Cy3 fluorescence is shown in blue, while Cy5 fluorescence is shown in red. **e,** Zoomed-in trajectories showing FRET transitions for 25-s segments in **d**. **f,** Transition occupancy density plots (TODPs) illustrating the most common FRET transitions. **g,** Box-and-whisker plot of stepping kinetics in the high- and low-FRET states for ***$W_{6\_13\_6}$*** on ***3-Foothold*** DNA Origami.

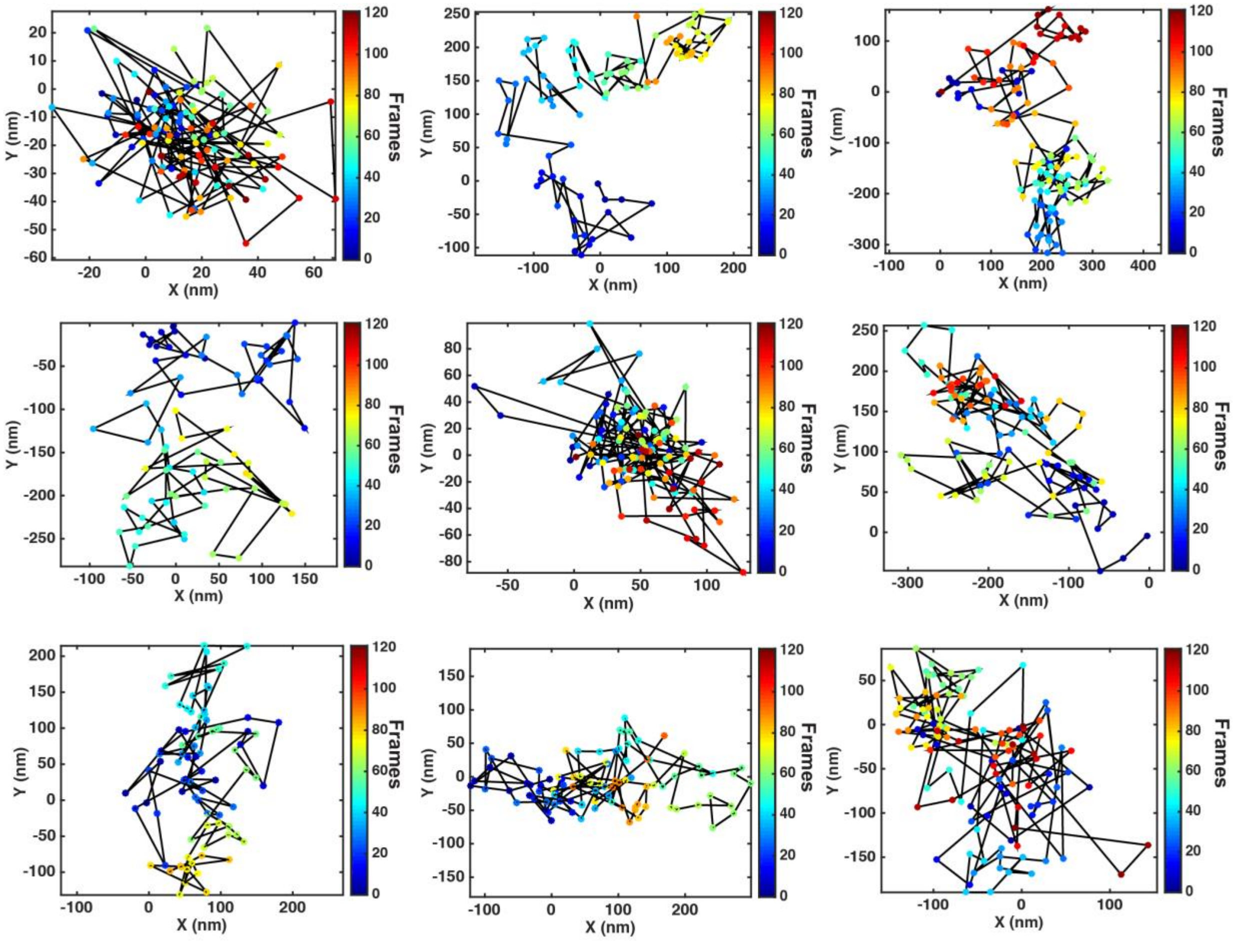


Figure A1. 10. Representative 2D particle tracking trajectories of ***$W_{6\_13\_6}$*** on surface coated with ***F_1*** and ***F_2***. One frame was acquired every 30 s.

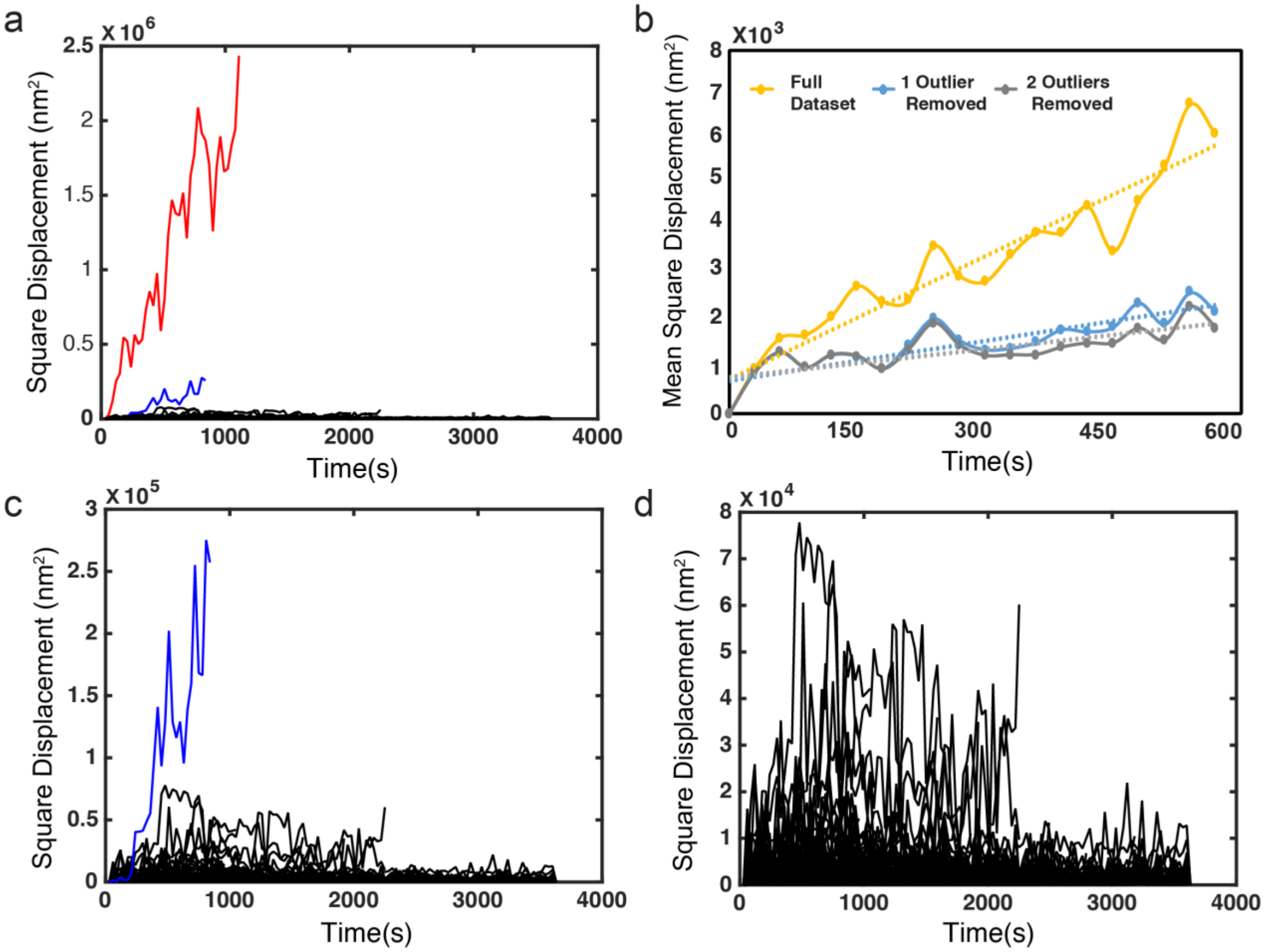


Figure A1. 11. MSD comparison for ***W**$_{8\_13\_8}$*. **a,** Square displacement for all trajectories (353 molecules). The extremely fast-moving outlier trajectory is highlighted in red. **b,** MSD comparison between all trajectories (n = 353 trajectories, yellow line), without the single fast-moving trajectory (n = 352 trajectories, blue line) and without the second fastest moving trajectory (n=351 trajectories, grey line). Dotted lines indicate linear regression fits to the data, resulting in calculated 2D diffusion coefficient estimates of 2.2 nm^2/s (yellow line), 0.7 nm^2/s (blue line), and 0.5 nm^2/s (grey line). Thus, removal of one very unusual out of 352 trajectories reduces the apparent diffusion coefficient >3-fold, and the second fastest trajectory shouldn't be count as an outlier. **c,** Square displacement for fastest moving outlier (red in a) removed trajectories. **d,** Square displacement for second fastest moving outlier (blue in a and c) removed trajectories.

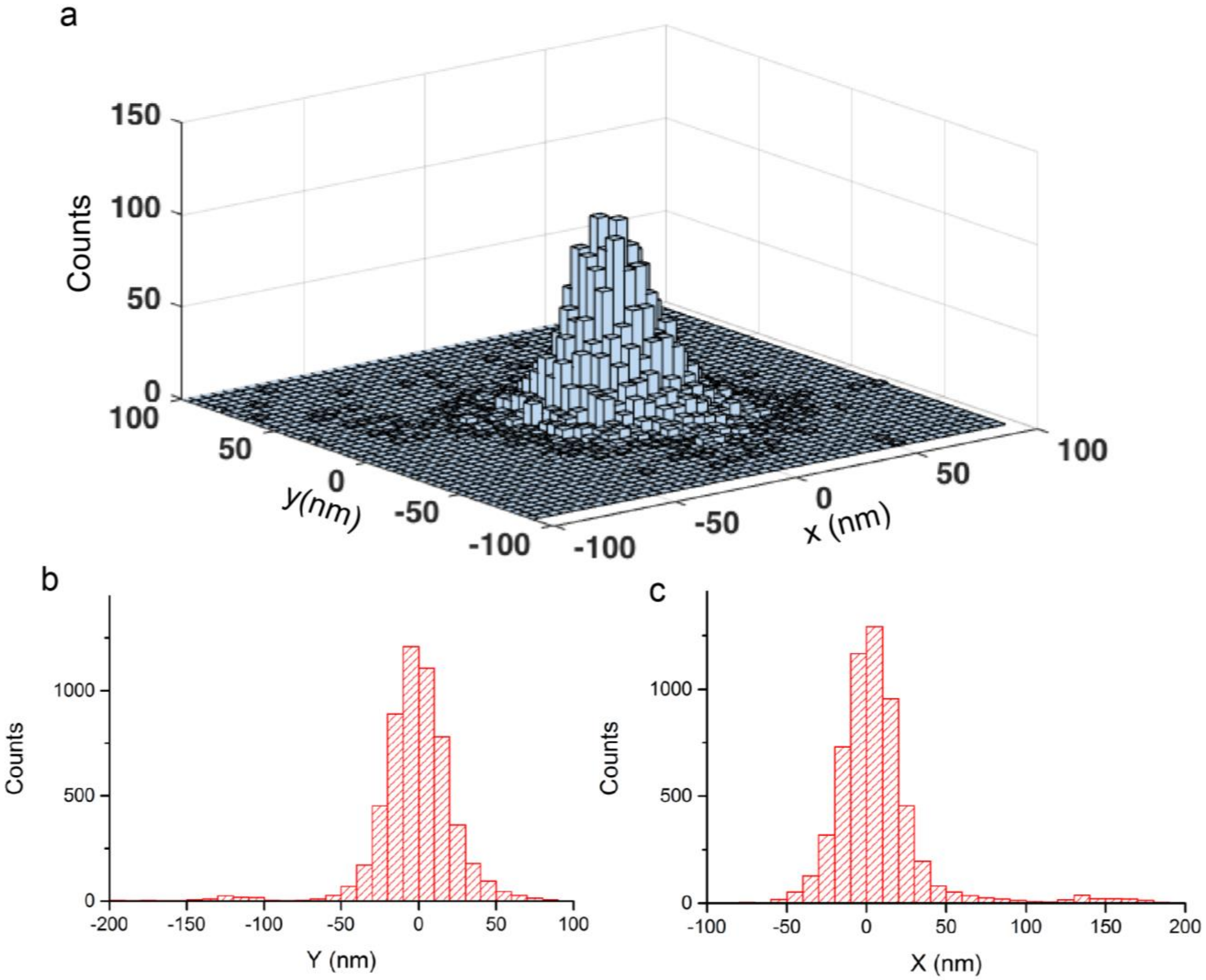


Figure A1. 12. Position distribution of ***$W_{6\_13\_6}$*** on a surface bearing only one foothold type (***F_1***) during 10 min of single-particle tracking. No walking is expected to occur on this control surface. **a,** 2D histogram showing the distribution of apparent x-y positions of all walkers (n=107) relative to their starting positions (0,0) over 10 min of observation. **b, c,** Histograms of walker coordinates in the x-(**b**) and y-(**c**)directions. The standard deviations of these coordinates (σ_x = 16.4 nm, σ_y = 15.2 nm) represent the approximate precision of localization in particle tracking experiments.

Table A1. 1. Staple sequences for the *3-Foothold* DNA origami design

Name	***Sequence (5′→3′)***
Oligo0	AGGTTTAGTACCGCCATGAGTTTCGTCACCAGTTTTCCAATC
Oligo1	GTATAAACAGTTAATGTGCGAATAATAATTTTTTTTCCAATC
Oligo2	CAGGAGGTTGAGGCAGAGGGAGTTAAAGGCCGTTTTCCAATC
Oligo3	TTCATCGGCATTTTCGTACACTAAAACACTCATTTTCCAATC
Oligo4	TTATTCATTAAAGGTGATGAACGGTGTACAGATTTTCCAATC
Oligo5	TACGCAGTATGTTAGCTCATTGTGAATTACCTTTTTCCAATC
Oligo6	GATAACCCACAAGAATGAGGCATAGTAAGAGCTTTTCCAATC
Oligo7	GTTACAAAATAAACAGAGTTCAGAAAACGAGATTTTCCAATC
Oligo8	TAGCAAGCAAATCAGATACCTTTAATTGCTCCTTTTCCAATC
Oligo9	ATCAACAATAGATAAGCATTTCGCAAATGGTCTTTTCCAATC
Oligo10	TAAAGCCAACGCTCAATTATGACCCTGTAATATTTTCCAATC
Oligo11	CAAGACAAAGAACGCGAATGCCGGAGAGGGTATTTTCCAATC
Oligo12	CTGTAAATCGTCGCTATAAACGTTAATATTTTTTTTCCAATC
Oligo13	ATTGCTTTGAATACCATGGGATAGGTCACGTTTTTTCCAATC
Oligo14	TTCATCAATATAATCCGTGCGGGCCTCTTCGCTTTTCCAATC
Oligo15	TCAATAGATAATACATTGGCTAGTACCCGTATTTTTCCAATC
Oligo16	CACCGCCTGCAACAGTCCGCTTTCCAGTCGGGTTTTCCAATC
Oligo17	AGGGACATTCTGGCCACAGCAGGCGAAAATCCTTTTCCAATC
Oligo18	TACAAACTACAACGCCTATCACCGTACTCAGGTTATCCATTC
Oligo19	TTCACGTTGAAAATCTTTGAGTAACAGTGCCCTTATCCATTC
Oligo20	CTTTTGCGGGATCGTCCCGCCGCCAGCATTGATTATCCATTC
Oligo21	TCTTTGACCCCCAGCGCAGACTGTAGCGCGTTTTATCCATTC
Oligo22	CCAGGCGCATAGGCTGTAAATATTGACGGAAATTATCCATTC
Oligo23	TATGCGATTTTAAGAAGATTAAGACTCCTTATTTATCCATTC
Oligo24	AACACTATCATAACCCGCGCTAATATCAGAGATTATCCATTC
Oligo25	ATGACCATAAATCAAAAGAGCCTAATTTGCCATTATCCATTC
Oligo26	TTTTGATAAGAGGTCATCATTACCGCGCCCAATTATCCATTC
Oligo27	AATAACCTGTTTAGCTCAGAACGCGCCTGTTTTTATCCATTC
Oligo28	CTTTTGCGGGAGAAGCCAAATTCTTACCAGTATTATCCATTC
Oligo29	GCTATTTTTGAGAGATGATGCAAATCCAATCGTTATCCATTC
Oligo30	GTTAAAATTCGCATTAGTGAATAACCTTGCTTTTATCCATTC
Oligo31	GGTGTAGATGGGCGCAATAACGGATTCGCCTGTTATCCATTC
Oligo32	TATTACGCCAGCTGGCTATCAGATGATGGCAATTATCCATTC
Oligo33	AAGGATCCCCGGGTACTAATAGATTAGAGCCGTTATCCATTC
Oligo34	AAACCTGTCGTGCCAGAGGCGGTCAGTATTAATTATCCATTC
Oligo35	TGTTTGATGGTGGTTCCACGACCAGTAATAAATTATCCATTC
Oligo36	CCCTCAGAACCGCCACAAGCCCAATAGGAACCTTTTCATACC
Oligo37	GGAACCTATTATTCTGAGTGAGAATAGAAAGGTTTTCATACC
Oligo38	TTGATATTCACAAACAATAACCGATATATTCGTTTTCATACC
Oligo39	TTAGCGTTTGCCATCTGCACCAACCTAAAACGTTTTCATACC
Oligo40	CGACTTGAGCCATTTGAACCGAACTGACCAACTTTTCATACC
Oligo41	ATACATAAAGGTGGCATGGGCTTGAGATGGTTTTTTCATACC
Oligo42	AATAAGAGCAAGAAACAATGCAGATACATAACTTTTCATACC
Oligo43	CAATCCAAATAAGAAAATTCATTGAATCCCCCTTTTCATACC
Oligo44	CGGTATTCTAAGAACGAAGCAAACTCCAACAGTTTTCATACC
Oligo45	ATAATATCCCATCCTAGAACGAGTAGATTTAGTTTTCATACC
Oligo46	GAGAATCGCCATATTTGCATAAAGCTAAATCGTTTTCATACC
Oligo47	ATATATTTTAGTTAATTATGATATTCAACCGTTTTTCATACC
Oligo48	TTAGAATCCTTGAAAAAGGAAGATTGTATAAGTTTTCATACC
Oligo49	CAGAGGCGAATTATTCTCCGTGGGAACAAACGTTTTCATACC
Oligo50	TACTTCTGAATAATGGCAGGCTGCGCAACTGTTTTTCATACC
Oligo51	TATTAGACTTTACAAACGAGGCAAGTCCGCTATTTTCATACC
Oligo52	AGCAGCAAATGAAAAATAACTCACATTAATTGTTTTCATACC
Oligo53	TTCTGACCTGAAAGCGGTTGCAGCAAGCGGTCTTTTCATACC
Oligo54	CATGTACCGTAACACCCCTCAGAACCGCCATTTTCATCAC
Oligo55	AACAACTAAAGGAATCCCCCTGCCTATTTCTTTTCATCAC
Oligo56	GTCGCTGAGGCTTGCGTCAGACGATTGGCCTTTTCATCAC
Oligo57	AAAGAGGCAAAAGAAGTCATAGCCCCCTTATTTTCATCAC
Oligo58	TTTGAAAGAGGACAGAATTATCACCGTCACTTTTCATCAC

Oligo59	TAATTTCAACTTTAAAAACGTAGAAAATACTTTTCATCAC
Oligo60	GCCAAAAGGAATTACTGAGTTAAGCCCAATTTTTCATCAC
Oligo61	TCAAATGCTTTAAACCCATATTATTTATCCTTTTCATCAC
Oligo62	GTCAGGATTAGAGAGTATAGAAGGCTTATCTTTTCATCAC
Oligo63	TTTGACCATTAGATATCCTGAACAAGAAAATTTTCATCAC
Oligo64	GTTGTACCAAAAACACAGTAGGGCTTAATTTTTTCATCAC
Oligo65	TCTAGCTGATAAATTAGAAAACTTTTTCAATTTTCATCAC
Oligo66	CAAATATTTAAATTGTTAATTAATTTTCCCTTTTCATCAC
Oligo67	GCGGATTGACCGTAAAGTTACAAAATCGCGTTTTCATCAC
Oligo68	TGGGAAGGGCGATCGTGATTGTTTGGATTATTTTCATCAC
Oligo69	GCGACCGTATACGCATTGAGGATTTAGAAGTTTTCATCAC
Oligo70	CGTTGCGCTCACTGCGCCACGCTGAGAGCCTTTTCATCAC
Oligo71	CACGCTGGTTTGCCCACAGAGATAGAACCCTTTTCATCAC
Oligo72	ATAAGTGCCGTCGAGAGCGTAACGATCTAAAGTTTTCTACAC
Oligo73	GATGATACAGGAGTGTTTGTATCGGTTTATCATTTTCTACAC
Oligo74	CCTCAGAGCCACCACCAGGGTAGCAACGGCTATTTTCTACAC
Oligo75	AGCAGCACCGTAATCAAGATTTGTATCATCGCTTTTCTACAC
Oligo76	GCGCCAAAGACAAAAGAACCGGATATTCATTATTTTCTACAC
Oligo77	AAACCGAGGAAACGCAAGAAAAATCTACGTTATTTTCTACAC
Oligo78	ACGGGAGAATTAACTGAGCGAGAGGCTTTTGCTTTTCTACAC
Oligo79	GCTACAATTTTATCCTGAAGCAAAGCGGATTGTTTTCTACAC
Oligo80	CGCACTCATCGAGAACAATATAATGCTGTAGCTTTTCTACAC
Oligo81	AGTAATTCTGTCCAGATGGCATCAATTCTACTTTTTCTACAC
Oligo82	GAATCATAATTACTAGGAACCCTCATATATTTTTTTCTACAC
Oligo83	TTTTAACCTCCGGCTTTCTGGAGCAAACAAGATTTTCTACAC
Oligo84	TGAATTACCTTTTTTAAATAGGAACGCCATCATTTTCTACAC
Oligo85	CGTCAGATGAATATACCGACGACAGTATCGGCTTTTCTACAC
Oligo86	AACAAAGAAACCACCATGGGTAACGCCAGGGTTTTTCTACAC
Oligo87	AGGAATTGAGGAAGGTGTTTCCTGTGTGAAATTTTTCTACAC
Oligo88	CATTAAAAATACCGAAGAGGCGGTTTGCGTATTTTTCTACAC
Oligo89	CTCAATCGTCTGAAATGAATAGCCCGAGATAGTTTTCTACAC
Oligo90	TTTTGTCGTCTTTCCCTCAGTACCAGGCGGTTATCTTCCA
Oligo91	GCTTGCTTTCGAGGTTCATACATGGCTTTTTTATCTTCCA
Oligo92	CAGAGGCTTTGAGGACCTCAGAACCGCCACTTATCTTCCA
Oligo93	CTGATAAATTGTGTCAATGAAACCATCGATTTATCTTCCA
Oligo94	CCCAAATCAACGTAATCATATGGTTTACCATTATCTTCCA
Oligo95	ATAAAACGAACTAACAAAGTTACCAGAAGGTTATCTTCCA
Oligo96	AAAAGAAGTTTTGCCAGGGAAGCGCATTAGTTATCTTCCA
Oligo97	CATCAAAAAGATTAAGCTATTTTGCACCCATTATCTTCCA
Oligo98	TCAACATGTTTTAAATATTAAACCAAGTACTTATCTTCCA
Oligo99	AATAGTAGTAGCATTACCGACAAAAGGTAATTATCTTCCA
Oligo100	TAAATGCAATGCCTGTAAGAATAAACACCGTTATCTTCCA
Oligo101	GAATCGATGAACGGTTCTGAGAGACTACCTTTATCTTCCA
Oligo102	AAAATAATTCGCGTCTTTAACAATTTCATTTTATCTTCCA
Oligo103	CTCAGGAAGATCGCAGATTTTCAGGTTTAATTATCTTCCA
Oligo104	TTTCCCAGTCACGACATTATCATTTTGCGGTTATCTTCCA
Oligo105	TGTTATCCGCTCACAAAATCAACAGTTGAATTATCTTCCA
Oligo106	TGGGCGCCAGGGTGGGCCCTAAAACATCGCTTATCTTCCA
Oligo107	GGTTGAGTGTTGTTCACCTACATTTTGACGTTATCTTCCA
Oligo108	GATTAGCGGGGTTTTGAGACGTTAGTAAATGATTTTACCCAT
Oligo109	ACCGTTCCAGTAAGCGGAATTTCTTAAACAGCTTTTACCCAT
Oligo110	CCCTCAGAGCCGCCACCTAAAGACTTTTTCATTTTTACCCAT
Oligo111	GGCCGGAAACGTCACCGAAATCCGCGACCTGCTTTTACCCAT
Oligo112	ACAATCAATAGAAAATCAAAGCTGCTCATTCATTTTACCCAT
Oligo113	AAGCAGATAGCCGAACGGAACAACATTATTACTTTTACCCAT
Oligo114	AGAATAACATAAAAACAGAGGGGGTAATAGTATTTTACCCAT
Oligo115	TAAATCAAGATTAGTTGAGGAAGCCCGAAAGATTTTACCCAT
Oligo116	TCATTCCAAGAACGGGTATGCAACTAAAGTACTTTTACCCAT
Oligo117	TAAGAGAATATAAAGTAACATCCAATAAATCATTTTACCCAT
Oligo118	TAAATAAGGCGTTAAAAGTAATGTGTAGGTAATTTTACCCAT
Oligo119	TTATCAAAATCATAGGAATCGTAAAACTAGCATTTTACCCAT
Oligo120	AACAAAATTAATTACATGGCCTTCCTGTAGCCTTTTACCCAT

Oligo121	ATAAAGAAATTGCGTACTCCAGCCAGCTTTCCTTTTACCCAT
Oligo122	TAAAAGTTTGAGTAACGTTGTAAAACGACGGCTTTTACCCAT
Oligo123	TATCTGGTCAGTTGGCATTCCACACAACATACTTTTACCCAT
Oligo124	AATGCGCGAACTGATATTTTTCTTTTCACCAGTTTTACCCAT
Oligo125	AAACGCTCATGGAAATCAGTTTGGAACAAGAGTTTTACCCAT
Oligo126	ATTTTCTGTATGGGATTCAAGAGAAGGATTAGTTTACTCACT
Oligo127	TTGATACCGATAGTTGCGCAGTCTCTGAATTTTTTACTCACT
Oligo128	GAGGAAGTTTCCATTAACCACCGGAACCGCCTTTTACTCACT
Oligo129	TCCATGTTACTTAGCCCCATTACCATTAGCAATTTACTCACT
Oligo130	GTGAATAAGGCTTGCCATAAGTTTATTTTGTCTTTACTCACT
Oligo131	AGGTAGAAAGATTCATCCTTTTTAAGAAAAGTTTTACTCACT
Oligo132	AAATGTTTAGACTGGAAGCAGCCTTTACAGAGTTTACTCACT
Oligo133	CTTCAAATATCGCGTTGGAGGTTTTGAAGCCTTTTACTCACT
Oligo134	GGTGTCTGGAAGTTTCCGGCTGTCTTTCCTTATTTACTCACT
Oligo135	TACAGGCAAGGCAAAGATTTTCGAGCCAGTAATTTACTCACT
Oligo136	AGATTCAAAAGGGTGAAATACCGACCGTGTGATTTACTCACT
Oligo137	TGTCAATCATATGTACAGAGTCAATAGTGAATTTTACTCACT
Oligo138	AGCTTTCATCAACATTAAACAAACATCAAGAATTTACTCACT
Oligo139	GGCACCGCTTCTGGTGTGCACGTAAAACAGAATTTACTCACT
Oligo140	CAGTGCCAAGCTTGCACGAACGTTATTAATTTTTTACTCACT
Oligo141	GAGCCGGAAGCATAAATCAAACCCTCAATCAATTTACTCACT
Oligo142	TGAGACGGGCAACAGCATGGCTATTAGTCTTTTTTACTCACT
Oligo143	TCCACTATTAAAGAACGCCATTGCAACAGGAATTTACTCACT
Oligo144	AGAGGCTGAGACTCCTTTGCTAAACAACTT
Oligo145	AGCCAGAATGGAAAGCGCCGACAATGACAA
Oligo146	ACCGGAACCAGAGCCAACGGGTAAAATACG
Oligo147	AAATCACCAGTAGCAGGAACGAGGCGCAGA
Oligo148	CAAAGACACCACGGACTGACGAGAAACACC
Oligo149	CTATCTTACCGAAGCCAGTTGAGATTTAGG
Oligo150	GTCAAAAATGAAAATTAGCGTCCAATACTG
Oligo151	ACCTCCCGACTTGCGTTAATTCGAGCTTCA
Oligo152	AAACCAATCAATAATATTCCATATAACAGT
Oligo153	AATTTAGGCAGAGGCAATTAGCAAAATTAA
Oligo154	AATTTAATGGTTTGAGAAAGGCCGGAGACA
Oligo155	ATTAAGACGCTGAGACCCGGTTGATAATCA
Oligo156	CAAAAGAAGATGATGAAATGTGAGCGAGTA
Oligo157	CATATCAAAATTATTCCGGAAACCAGGCAA
Oligo158	ATTAAATCCTTTGCCTGCCTGCAGGTCGAC
Oligo159	GCTGAACCTCAAATAGTGTAAAGCCTGGGG
Oligo160	AGACAATATTTTTGATGATTGCCCTTCACC
Oligo161	ACAATATTACCGCCAGTGGACTCATATCCA
Oligo162	CATTTTCAGGGATAGCCCTCAGAGCCACCACC
Oligo163	TCAACAGTTTCAGCGGAAACATGAAAGTATTA
Oligo164	CAACCATCGCCCACGCAATAAATCCTCATTAA
Oligo165	TAATGCCACTACGAAGTTTCATAATCAAAATC
Oligo166	CGGTCAATCATAAGGGGGAATTAGAGCCAGCA
Oligo167	AGAACGAGTAGTAAATACATATAAAAGAAACG
Oligo168	AATACCACATTCAACTAATGAAATAGCAATAG
Oligo169	CGGAATCGTCATAAATCGATTTTTTGTTTAAC
Oligo170	AAGCGAACCAGACCGGCGAGGCGTTTTAGCGA
Oligo171	TGATTCCCAATTCTGCATTTACGAGCATGTAG
Oligo172	GCAATAAAGCCTCAGAAACAACGCCAACATGT
Oligo173	GTCAAATCACCATCAATTCATCTTCTGACCTA
Oligo174	GAAAAGCCCCAAAAACCATAGCGATAGCTTAG
Oligo175	ACAACCCGTCGGATTCATTTCAATTACCTGAG
Oligo176	AGCGCCATTCGCCATTAAGGGTTAGAACCTAC
Oligo177	TCTAGACCTTTGATAGCAATTCGACAACTCGT
Oligo178	TGCCTAATGAGTGAGCTCTAAAGCATCACCTT
Oligo179	GCCTGGCCCTGAGAGATAAGAATACGTGGCAC
Oligo180	AGCCCGGAATAGGTGTGTAGCATTCCACAGTTTCACTACT
Oligo181	AACGGGGTCAGTGCCCCAAAAAAAAGGCTCTTTCACTACT
Oligo182	GAACCACCACCAGAGACCCTCAGCAGCGAATTTCACTACT

Oligo183	GTTTGCCTTTAGCGTATTATACCAAGCGCGTTTCACTACT
Oligo184	TTGAGGGAGGGAAGGGCTGACCTTCATCAATTTCACTACT
Oligo185	CAAAAGAACTGGCATCTGGCTCATTATACCTTTCACTACT
Oligo186	TCAGAGGGTAATTGATCGTTTACCAGACGATTTCACTACT
Oligo187	AACGAGCGTCTTTCCAATCAGGTCTTTACCTTTCACTACT
Oligo188	TTTTCATCGTAGGAATTTTTGCGGATGGCTTTTCACTACT
Oligo189	CATGTTCAGCTAATGATATTTTCATTTGGGTTTCACTACT
Oligo190	ATCATATGCGTTATACTTTATTTCAACGCATTTCACTACT
Oligo191	CTATATGTAAATGCTCTACAAAGGCTATCATTTCACTACT
Oligo192	AATCAATATATGTGAAATTTTTGTTAAATCTTTCACTACT
Oligo193	ACATCGGGAGAAACATCGTAACCGTGCATCTTTCACTACT
Oligo194	CATCATATTCCTGATGAAAGGGGGATGTGCTTTCACTACT
Oligo195	AGGAGCACTAACAACCGAGCTCGAATTCGTTTTCACTACT
Oligo196	GATAAAACAGAGGTGCTGCATTAATGAATCTTTCACTACT
Oligo197	CAGATTCACCAGTCACGAAATCGGCAAAATTTTCACTACT
Oligo198	ACAGCCCTCATAGTTAGGGTTGATATAAGTATTTTTTAACCC
Oligo199	CAAAAGGAGCCTTTAAACTGGTAATAAGTTTTTTTTTAACCC
Oligo200	AGACAGCATCGGAACGCTCAGAGCCGCCACCATTTTTAACCC
Oligo201	AAACAAAGTACAACGGGTAGCGACAGAATCAATTTTTAACCC
Oligo202	GAGTAATCTTGACAAGGGCGACATTCAACCGATTTTTAACCC
Oligo203	AGTCAGGACGTTGGGAATAATAACGGAATACCTTTTTAACCC
Oligo204	CGATAAAAACCAAAATAACACCCTGAACAAAGTTTTTAACCC
Oligo205	CTGACTATTATAGTCAGAATCTTACCAACGCTTTTTTAACCC
Oligo206	TAGAGCTTAATTGCTGAAGCAAGCCGTTTTTATTTTTAACCC
Oligo207	GCGCGAGCTGAAAAGGCGACGACAATAAACAATTTTTAACCC
Oligo208	AGGATAAAAATTTTTAAAAAAGCCTGTTTAGTTTTTTAACCC
Oligo209	GGTCATTGCCTGAGAGAGGTTGGGTTATATAATTTTTAACCC
Oligo210	AGCTCATTTTTTAACCATGGAAACAGTACATATTTTTAACCC
Oligo211	TGCCAGTTTGAGGGGAAGTAACAGTACCTTTTTTTTTAACCC
Oligo212	TGCAAGGCGATTAAGTGAAGGAGCGGAATTATTTTTTAACCC
Oligo213	AATCATGGTCATAGCTTATCTAAAATATCTTTTTTTTAACCC
Oligo214	GGCCAACGCGCGGGGACGAACCACCAGCAGAATTTTTAACCC
Oligo215	CCCTTATAAATCAAAAGGATTATTTACATTGGTTTTTAACCC
Oligo228	ATTACGCCTGAGGGGACGACGACAGGAACAAAGGTGACTGCTTCTAC
Oligo229	GGGAAGGGAGATCGCACTCCAGCCGAGCGAGTGGGACGCTCATTTTCA
Oligo230	CGCCATTTTCTGGTGCCGGAAACCTGTAGCACAAGACCATGCTTTG
Oligo231	ACTAGCATAGCCCCAAAAACAGGAAACGCCATCCATCGTTTTCTATC
Oligo232	AACAAGAGATATTTAAATTGTAAATGTTAAATTCGGGACAAGTCTCTC
Oligo233	CCTGTGTGTACGAGCCGGAAGCATGTTTTTCT
Oligo234	ATGGTCATACGACGTTGTAAAACGTCTTCGCTACGACGGCCCCTAAT
Oligo235	CTCGAATGGTGCCTAATGAGTGAGAGGCGG
Oligo236	ATCCCCGGCTTGCATGCCTGCAGGCAACTGTTAGCCTGCACAGACAGC
Oligo237	TAGTACCCTTGCGTTGCGCTCACTAGCTGCAT
Oligo238	CCGTATAGATAGCGAGGCAAGTAGGCAAAGACTACATGTATCTCGA
Oligo239	AATCACCAAAAAACATTATGACCCAGCTAAAT
Oligo240	AGGCCGGAGTTCTAGCTGATAAATATCGTAAAGAGAGTGACAGATGT
Oligo241	TCAAAAGGAGAAGCCTTTATTTCAAAATTA
Oligo242	ATGTGTAGTAGCTATTTTTGAGAGCTGGAGCAACCTGGCCTGCGTATC
Oligo243	ACACAACAAAATTGTTTCCACATACGACAAAC
Oligo244	CCAGGGTGAAAGTGTAAAGCCTGGTCGTAATCCTGCTTCCCTACGCT
Oligo245	TTTGCGTATTGGGCGGTTGCAGCAAGCGGTGTTGAGTGTTGTTCC
Oligo246	GCGCGGGGAGCTAACTCACATTAAGTATAAGGCAAATTCAGATGACTC
Oligo247	TAATGAATCGGCCAACCAGCAGGCGAAAATCCCCTTATAAATCAAAAG
Oligo248	TCGTGCCGCCCGCTTTCCAGTCGCTAGCGAGTGCAGAAAGGCTGTC
Oligo249	CGGTTGTAGAAACCTGCAAATGGTCAATAACCGAAGGCACATACATTT
Oligo250	GAGCATAATGTAATACTTTTGCGGGGTGAGAAGCCGCCCAACTGAGG
Oligo251	AGCAATAAAGCCTCACATTTGGGGCGCGAGATTAAACGGGTAAAA
Oligo252	AGAATTAGCAACGCAAGGATAAAACCTGAGTAGGTGCATAAACGCAAC
Oligo253	CATACAGGCAAGGCAAAATTCTACTAATAGTAAGGACTAAAGACTTTT
Oligo254	CTGATTGCCCTTCACCCCACTATTAAAGAACG
Oligo255	TAACGCCATATCATAACCCTCGTTAAAACGAGAGGTCTGGACGCTACA
Oligo256	AACTAATGCAGATACACTCCAACTTATGTGTACGGCGGATTGACCGTA

Oligo257	AGGAATAAAACCAAAATAGCGAGAATCCCCTCGATGTTAGTTCGTC
Oligo258	TCATCAGTTGAGATTTAAGAGTTGTGGACTAGAACAACCCGTCGGATT
Oligo259	TTACAGGTAAGTTTTGCCAGAGGGCCAATACTGGATACTCTTGGTTC
Oligo260	AACGGAACAACATTACCCGCTTGATATGAACAGCTTTCATCAACA
Oligo261	GTTAATAATGAATAAGGCTTGCCCACAAAGCTCCATGGGCGTCCCTAC
Oligo262	GGGAAGAAAAATCTACACCTGTGCGGAGCAAGCAAAAATAATTCGCGT
Oligo263	ACCAGTCAACGAGTAGTAAATTGAACCGGATAGTGGTGATGGCAGA
Oligo264	AGAACTGGCTCATTATTCCCAGGACCACGATTCAGCTCATTTTTTAAC
Oligo265	AATGACCATATAGTCAGAAGCAAAATGGCTTAAAGCCTGGGTTAAAAA
Oligo266	CAGTTCAGTACCAGACGACGATAACCACATTC
Oligo267	CTCAAATAGATTAAGAGGAAGCTGCTCCTTTCAATTCTGTAGCACG
Oligo268	TATTCATTGAGGCTTTTGCAAAAGAGAAAGAT
Oligo269	GCGGAATCATCGCGTTTTAATTCGCCAACAGGGGCGGATGATTAGTG
Oligo270	ATAGCGTGGTAATAGTAAAATGAACGAACT
Oligo271	GCTCATTCCAGACGGTCAATCATACTTAGCCGTTGATTATGGAATCGA
Oligo272	TCAACGTATGACGAGAAACACCAGAGGACGTT
Oligo273	TATTCATCAACTTTGAAAGAGGTGTGTCGAGCTACGTCAATGAACC
Oligo274	CTTGACAAGGGCTTGAGATGGTTTCGATTTTA
Oligo275	GAGCTTAATTAAATATGCAACTAAACAAGAGTGCCTGGCCCTGAGAGA
Oligo276	TTGATAAGTTTCATTCCATATAGAGATAGGCCACGCTGGTTTGCCC
Oligo277	TCAGGATTTCTGCGAACGAGTAGAGCAAAATCTGTTTGATGGTGGTT
Oligo278	GAACGAGGACCTAAAACGAAAGAGGCCACTACTGTTTAGCTATATTTT
Oligo279	AATCCGCAAACACTCATCTTTGAAGTTTCCCTGAAAAGGTGGCATC
Oligo312	ATGGGATACGTGCATCTGCCAGTTAGCTGGCG
Oligo313	CTCCGTGGTATCGGCCTCAGGACGATCGGT
Oligo314_T2	TTAAATGTAGCTTTCCGGCACCGCCGCCATTCttttttCCCTCATTCAATACCCCTACG
Oligo315	CTGGCCTTGGTTGATAATCAGAAAGTCAATCA
Oligo316	CAATAGGAGATTGTATAAGCAAAATCGATG
Oligo317	AAAGGGGGCAGGGTTTTCCCAGTCAGCTGTTT
Oligo318_T1'	CAATACCCCTACGGTCACTTCttttttGCGGGCCACGGCCAGTGCCAAGGTACCGAG
Oligo319	AGGCTGCGTCGACTCTAGACCTTTCGCATGGC
Oligo320_T1_Cy3	/5Cy3/CAATACCCCTACGGTCACTTCttttttTATGTACCAATATGATATTCAACCGACAGTCA
Oligo321	AACGGTATAATGCCGGAGAGGGGTAAAGAT
Oligo324	GGAAGCAGTTTTTAACCCAGGCTTGTTTGTCGTATGTGGAACGGCCT
Oligo325	CATCCGCCGAGTCATCTGAATTTGCGTGCTACAGAATTGAAGCGTAG
Oligo326	TGGGCGGCTCGATTCCATAATCAAGACAGCCTTTCTGCACCACTAAT
Oligo327	TACTCAAAGTTGCGTTTATGCACCGGTTCATTGACGTAGCCCTCAGT
Oligo328	GCCGTCGTTGTAGCGTCCAGACCTCCAACCGGTATAGGAAAGTTAAT
Oligo329	GAGTATCCGCTGTCTGTGCAGGCTGACGAACTAACATCGAATTAGGG
Oligo330	TCACTCTCGTAGGGACGCCCATGGTCGAGATACATGTAGTGAACCAA
Oligo331	AGCATAGAGATACGCAGGCCAGGTTCTGCCATCACCACTAACATCTG
Oligo332	CAGTCACCTACACATAAGTTGGAGACACCTAGGGAGCACGGCCATAC
Oligo333	CAAGCGGGTGAAAATGAGCGTCCCCTAGTCCACAACTCTTGTAGAAG
Oligo334	AACGATGGCTTGCTCCGCACAGGTCAAAGCATGGTCTTGTTTCATAT
Oligo335	AGGTCAGGGAGAGACTTGTCCCGAAATCGTGGTCCTGGGAGATAGAA
Oligo348	TTTTGCGGGCGGATTGCATCAAAAGCTTTAAATTTTGTGATGAA
Oligo349	CCTTTAATCCGAAAGACTTCAAATGTCATAAATTTTGAATGGAT
Oligo350	GCAAACTAGCTTCAAAGCGAACTAGACTGGTTTTGGGTTAAA
Oligo351	CCATGTTAAGGGAACCGAACTGACTACCCAAATTTTTGGAAGAT
Oligo352	TGATAAATACAGATGAACGGTGTAAGAGTAATTTTTAGTGAGTA
Oligo353	AGTTTGGAAGTACGGTGTCTGGAAGAGGTCATTTTTGGTATGAA
Oligo354	AATAGCCCACAGTTGATTCCCAATAGAGAGTATTTTGATTGGAA
Oligo355	GAAATCGTTTAGTTTGACCATTGACCGGAATTTTAGTAGTGA
Oligo356	TACGTAATGCAAAAGAATACACTAGACCTGCTTTTTGTGTAGAA
Oligo357	TCATGAGGACCCCCAGCGATTATATCATCGCCTTTTATGGGTAA
Oligo360	TTTTGCGGGCGGATTGCATCAAAAGCTTTAAA
Oligo361	CCTTTAATCCGAAAGACTTCAAATGTCATAAA
Oligo362	GCAAACTAGCTTCAAAGCGAACTAGACTGG
Oligo363	CCATGTTAAGGGAACCGAACTGACTACCCAAA
Oligo364	TGATAAATACAGATGAACGGTGTAAGAGTAAT
Oligo365	AGTTTGGAAGTACGGTGTCTGGAAGAGGTCATttttt/3bio/
Oligo366	AATAGCCCACAGTTGATTCCCAATAGAGAGTAttttt/3bio/

Oligo367	GAAATCGTTTAGTTTGACCATTGACCGGAAttttt/3bio/
Oligo368	TACGTAATGCAAAAGAATACACTAGACCTGCTttttt/3bio/
Oligo369	TCATGAGGACCCCCAGCGATTATATCATCGCCttttt/3bio/

References

1. Sharonov, A. & Hochstrasser, R. M. Wide-field subdiffraction imaging by accumulated binding of diffusing probes. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 18911–18916 (2006).
2. Bates, W. M., Huang, B., Dempsey, G. T. & Zhuang, X. Multicolor Super-resolution Imaging with Photo-switchable Fluorescent Probes. *Science* **317**, 1749–1753 (2007).
3. Hess, S. T., Girirajan, T. P. K. & Mason, M. D. Ultra-High Resolution Imaging by Fluorescence Photoactivation Localization Microscopy. *Biophys. J.* **91**, 4258–4272 (2006).
4. Hell, S. W. Far-Field Optical Nanoscopy. *Science* **316**, 1153–1158 (2007).
5. Hell, S. W. & Wichmann, J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Opt. Lett.* **19**, 780–782 (1994).
6. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5**, 507–516 (2008).
7. Ha, T. Single-molecule fluorescence resonance energy transfer. *Methods San Diego Calif* **25**, 78–86 (2001).
8. Sabanayagam, C. R., Eid, J. S. & Meller, A. High-throughput scanning confocal microscope for single molecule analysis. *Appl. Phys. Lett.* **84**, 1216–1218 (2004)

9. Kapanidis, A. N. *et al.* Alternating-laser excitation of single molecules. *Acc. Chem. Res.* **38**, 523–533 (2005).

10. Best, R. B. *et al.* Effect of flexibility and cis residues in single-molecule FRET studies of polyproline. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 18964–18969 (2007).

11. Pinheiro, A. V., Han, D., Shih, W. M. & Yan, H. Challenges and opportunities for structural DNA nanotechnology. *Nat. Nanotechnol.* **6**, 763–772 (2011).

12. Seeman, N. C. Nucleic acid junctions and lattices. *J. Theor. Biol.* **99**, 237–247 (1982).

13. Liu, D., Wang, M., Deng, Z., Walulu, R. & Mao, C. Tensegrity: Construction of Rigid DNA Triangles with Flexible Four-Arm DNA Junctions. *J. Am. Chem. Soc.* **126**, 2324–2325 (2004).

14. Zheng, J. *et al.* From Molecular to Macroscopic via the Rational Design of a Self-Assembled 3D DNA Crystal. *Nature* **461**, 74–77 (2009).

15. Kural, C. *et al.* Kinesin and Dynein Move a Peroxisome in Vivo: A Tug-of-War or Coordinated Movement? *Science* **308**, 1469–1472 (2005).

16. Helenius, J., Brouhard, G., Kalaidzidis, Y., Diez, S. & Howard, J. The depolymerizing kinesin MCAK uses lattice diffusion to rapidly target microtubule ends. *Nature* **441**, 115–119 (2006).

17. Lund, K. *et al.* Molecular robots guided by prescriptive landscapes. *Nature* **465**, 206–210 (2010).

18. Cha, T.-G. *et al.* A synthetic DNA motor that transports nanoparticles along carbon nanotubes. *Nat. Nanotechnol.* **9**, 39–43 (2014).

19. Omabegho, T., Sha, R. & Seeman, N. C. A Bipedal DNA Brownian Motor with Coordinated Legs. *Science* **324**, 67–71 (2009).
20. Sherman, W. B. & Seeman, N. C. A Precisely Controlled DNA Biped Walking Device. *Nano Lett.* **4**, 1203–1207 (2004).
21. Vale, R. D. & Milligan, R. A. The way things move: looking under the hood of molecular motor proteins. *Science* **288**, 88–95 (2000).
22. Tian, Y., He, Y., Chen, Y., Yin, P. & Mao, C. A DNAzyme That Walks Processively and Autonomously along a One-Dimensional Track. *Angew. Chem. Int. Ed.* **44**, 4355–4358 (2005).
23. Wickham, S. F. J. *et al.* A DNA-based molecular motor that can navigate a network of tracks. *Nat. Nanotechnol.* **7**, 169–173 (2012).
24. Sherman, W. B. & Seeman, N. C. A Precisely Controlled DNA Biped Walking Device. *Nano Lett.* **4**, 1203–1207 (2004).
25. Li, X. & Liu, D. R. DNA-Templated Organic Synthesis: Nature’s Strategy for Controlling Chemical Reactivity Applied to Synthetic Molecules. *Angew. Chem. Int. Ed.* **43**, 4848–4870 (2004).
26. C. Dreaden, E., M. Alkilany, A., Huang, X., J. Murphy, C. & A. El-Sayed, M. The golden age: gold nanoparticles for biomedicine. *Chem. Soc. Rev.* **41**, 2740–2779 (2012).
27. King, S. J. & Schroer, T. A. Dynactin increases the processivity of the cytoplasmic dynein motor. *Nat. Cell Biol.* **2**, 20–24 (2000).
28. Zhang, D. Y. & Winfree, E. Control of DNA Strand Displacement Kinetics Using Toehold Exchange. *J. Am. Chem. Soc.* **131**, 17303–17314 (2009).

29. Fu, J. *et al.* Multi-enzyme complexes on DNA scaffolds capable of substrate channelling with an artificial swinging arm. *Nat. Nanotechnol.* **9**, 531–536 (2014).

30. Johnson-Buck, A. *et al.* Kinetic fingerprinting to identify and count single nucleic acids. *Nat. Biotechnol.* **33**, 730–732 (2015).

31. Hayward, S. L. *et al.* Ultraspecific and Amplification-Free Quantification of Mutant DNA by Single-Molecule Kinetic Fingerprinting. *J. Am. Chem. Soc.* **140**, 11755–11762 (2018).

32. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R*. (Springer-Verlag, 2013).

33. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. (Springer-Verlag, 2009).

34. Choi, T., Chan, H. K. & Yue, X. Recent Development in Big Data Analytics for Business Operations and Risk Management. *IEEE Trans. Cybern.* **47**, 81–92 (2017).

35. Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y. & Sun, X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decis. Support Syst.* **50**, 559–569 (2011).

36. Mullainathan, S. & Spiess, J. Machine Learning: An Applied Econometric Approach. *J. Econ. Perspect.* **31**, 87–106 (2017).

37. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).

38. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C-N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).

39. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**, 360–365 (2018).

40. Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).

41. Sainath, T. N. *et al.* Deep Convolutional Neural Networks for large-scale speech tasks. *Neural Netw. Off. J. Int. Neural Netw. Soc.* **64**, 39–48 (2015).

42. Hinton, G. *et al.* Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Process. Mag.* **29**, 82–97 (2012).

43. Mikolov, T., Deoras, A., Povey, D., Burget, L. & Cernocký, J. Strategies for training large scale neural network language models. *2011 IEEE Workshop Autom. Speech Recognit. Underst.* 196–201 (2011). doi:10.1109/ASRU.2011.6163930

44. Farabet, C., Couprie, C., Najman, L. & Lecun, Y. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1915–1929 (2013).

45. Tompson, J. J., Jain, A., LeCun, Y. & Bregler, C. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. in *Advances in Neural Information Processing Systems 27* (eds. Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q.) 1799–1807 (Curran Associates, Inc., 2014).

46. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing*

Systems 25 (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, Inc., 2012).

47. Leung, M. K. K., Xiong, H. Y., Lee, L. J. & Frey, B. J. Deep learning of the tissue-regulated splicing code. *Bioinforma. Oxf. Engl.* **30**, i121-129 (2014).
48. Xiong, H. Y. *et al.* RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
49. Shin, J.-S. & Pierce, N. A. A Synthetic DNA Walker for Molecular Transport. *J. Am. Chem. Soc.* **126**, 10834–10835 (2004).
50. Bath, J., Green, S. J. & Turberfield, A. J. A Free-Running DNA Motor Powered by a Nicking Enzyme. *Angew. Chem.* **117**, 4432–4435 (2005).
51. Omabegho, T., Sha, R. & Seeman, N. C. A Bipedal DNA Brownian Motor with Coordinated Legs. *Science* **324**, 67–71 (2009).
52. Gu, H., Chao, J., Xiao, S.-J. & Seeman, N. C. A Proximity-Based Programmable DNA Nanoscale Assembly Line. *Nature* **465**, 202–205 (2010).
53. He, Y. & Liu, D. R. Autonomous Multistep Organic Synthesis in a Single Isothermal Solution Mediated by a DNA Walker. *Nat. Nanotechnol.* **5**, 778–782 (2010).
54. Qian, L. & Winfree, E. Scaling Up Digital Circuit Computation with DNA Strand Displacement Cascades. *Science* **332**, 1196–1201 (2011).
55. Qian, L., Winfree, E. & Bruck, J. Neural network computation with DNA strand displacement cascades. *Nature* **475**, 368–372 (2011).
56. Douglas, S. M., Bachelet, I. & Church, G. M. A Logic-Gated Nanorobot for Targeted Transport of Molecular Payloads. *Science* **335**, 831–834 (2012).
57. Thubagere, A. J. *et al.* A cargo-sorting DNA robot. *Science* **357**, eaan6558 (2017).

58. Zhang, D. Y. & Seelig, G. Dynamic DNA nanotechnology using strand-displacement reactions. *Nat. Chem.* **3**, 103–113 (2011).

59. Zhang, D. Y., Chen, S. X. & Yin, P. Optimizing the specificity of nucleic acid hybridization. *Nat. Chem.* **4**, 208–214 (2012).

60. Wickham, S. F. J. *et al.* Direct observation of stepwise movement of a synthetic molecular transporter. *Nat. Nanotechnol.* **6**, 166–169 (2011).

61. Yehl, K. *et al.* High-speed DNA-based rolling motors powered by RNase H. *Nat. Nanotechnol.* **11**, 184–190 (2015).

62. Ha, T. Single-molecule fluorescence resonance energy transfer. *Methods San Diego Calif* **25**, 78–86 (2001).

63. Michelotti, N., de Silva, C., Johnson-Buck, A. E., Manzo, A. J. & Walter, N. G. A bird's eye view tracking slow nanometer-scale movements of single molecular nano-assemblies. *Methods Enzymol.* **475**, 121–148 (2010).

64. Aitken, C. E., Marshall, R. A. & Puglisi, J. D. An Oxygen Scavenging System for Improvement of Dye Stability in Single-Molecule Fluorescence Experiments. *Biophys. J.* **94**, 1826–1835 (2008).

65. Blanco, M. & Walter, N. G. Analysis of Complex Single-Molecule FRET Time Trajectories. *Methods Enzymol.* **472**, 153–178 (2010).

66. Nicolai, C. & Sachs, F. SOLVING ION CHANNEL KINETICS WITH THE QuB SOFTWARE. *Biophys. Rev. Lett.* **08**, 191–211 (2013).

67. Dill, K. A. & Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology*. (Garland Science, 2003).

68. Chen, H. *et al.* Ionic Strength-Dependent Persistence Lengths of Single-Stranded RNA and DNA. *Proc. Natl. Acad. Sci.* **109**, 799–804 (2012).

69. Particle Track and Analysis (PTA). Available at: http://www.sanken.osaka-u.ac.jp/labs/bse/ImageJcontents/frameImageJ.html. (Accessed: 29th August 2017)

70. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).

71. Pan, J., Li, F., Cha, T.-G., Chen, H. & Choi, J. H. Recent progress on DNA based walkers. *Curr. Opin. Biotechnol.* **34**, 56–64 (2015).

72. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5**, 507–516 (2008).

73. Walter, N. G., Huang, C.-Y., Manzo, A. J. & Sobhy, M. A. Do-it-yourself guide: how to use the modern single-molecule toolkit. *Nat. Methods* **5**, 475–489 (2008).

74. Srinivas, N. *et al.* On the biophysics and kinetics of toehold-mediated DNA strand displacement. *Nucleic Acids Res.* **41**, 10641–10658 (2013).

75. Panyutin, I. G. & Hsieh, P. Formation of a Single Base Mismatch Impedes Spontaneous DNA Branch Migration. *J. Mol. Biol.* **230**, 413–424 (1993).

76. Beattie, K. L., Wiegand, R. C. & Radding, C. M. Uptake of homologous single-stranded fragments by superhelical DNA. *J. Mol. Biol.* **116**, 783–803 (1977).

77. Jungmann, R. *et al.* Single-Molecule Kinetics and Super-Resolution Microscopy by Fluorescence Imaging of Transient Binding on DNA Origami. *Nano Lett.* **10**, 4756–4761 (2010).

78. Dupuis, N. F., Holmstrom, E. D. & Nesbitt, D. J. Single-Molecule Kinetics Reveal Cation-Promoted DNA Duplex Formation Through Ordering of Single-Stranded Helices. *Biophys. J.* **105**, 756–766 (2013).

79. Smith, S. B., Cui, Y. & Bustamante, C. Overstretching B-DNA: The Elastic Response of Individual Double-Stranded and Single-Stranded DNA Molecules. *Science* **271**, 795–799 (1996).

80. Jin, Z., Geißler, D., Qiu, X., Wegner, K. D. & Hildebrandt, N. A Rapid, Amplification-Free, and Sensitive Diagnostic Assay for Single-Step Multiplexed Fluorescence Detection of MicroRNA. *Angew. Chem. Int. Ed.* **54**, 10024–10029 (2015).

81. Kim, K., Oh, J.-W., Lee, Y. K., Son, J. & Nam, J.-M. Associating and Dissociating Nanodimer Analysis for Quantifying Ultrasmall Amounts of DNA. *Angew. Chem. Int. Ed.* **56**, 9877–9880 (2017).

82. Cohen, L., Hartman, M. R., Amardey-Wellington, A. & Walt, D. R. Digital direct detection of microRNAs using single molecule arrays. *Nucleic Acids Res.* **45**, e137–e137 (2017).

83. Zhang, D. Y., Chen, S. X. & Yin, P. Optimizing the specificity of nucleic acid hybridization. *Nat. Chem.* **4**, 208–214 (2012).

84. Kim, K., Oh, J.-W., Lee, Y. K., Son, J. & Nam, J.-M. Associating and Dissociating Nanodimer Analysis for Quantifying Ultrasmall Amounts of DNA. *Angew. Chem. Int. Ed Engl.* **56**, 9877–9880 (2017).

85. Zhang, D. Y., Chen, S. X. & Yin, P. Optimizing the specificity of nucleic acid hybridization. *Nat. Chem.* **4**, 208–214 (2012).

86. Johnson-Buck, A. *et al.* Kinetic fingerprinting to identify and count single nucleic acids. *Nat. Biotechnol.* **33**, 730–732 (2015).

87. Sahl, S. J., Hell, S. W. & Jakobs, S. Fluorescence nanoscopy in cell biology. *Nat. Rev. Mol. Cell Biol.* **18**, 685–701 (2017).

88. Wide Field of View | Yokogawa Electric Corporation. Available at: https://www.yokogawa.com/solutions/products-platforms/life-science/spinning-disk-confocal/csu-w1-confocal-scanner-unit/. (Accessed: 16th May 2018)

89. Home - Oxford Nanoimaging. Available at: https://www.oxfordni.com/. (Accessed: 16th May 2018)

90. Nucleic acid analogues. Available at: https://www.atdbio.com/content/12/Nucleic-acid-analogues. (Accessed: 16th May 2018)

91. TM Prediction Tool. Available at: https://www.exiqon.com/ls/Pages/ExiqonTMPredictionTool.aspx. (Accessed: 16th May 2018)

92. NUPACK: Analysis input. Available at: http://www.nupack.org/partition/new. (Accessed: 16th May 2018)

93. OligoAnalyzer 3.1 | IDT. Available at: https://www.idtdna.com/calc/analyzer. (Accessed: 16th May 2018)

94. Oligo Optimizer Tool. Available at: https://www.exiqon.com/ls/Pages/ExiqonOligoOptimizerTool.aspx. (Accessed: 16th May 2018)

95. Dupuis, N. F., Holmstrom, E. D. & Nesbitt, D. J. Single-Molecule Kinetics Reveal Cation-Promoted DNA Duplex Formation Through Ordering of Single-Stranded Helices. *Biophys. J.* **105**, 756–766 (2013).

96. Dupuis, N. F., Holmstrom, E. D. & Nesbitt, D. J. Single-Molecule Kinetics Reveal Cation-Promoted DNA Duplex Formation Through Ordering of Single-Stranded Helices. *Biophys. J.* **105**, 756–766 (2013).

97. Blake, R. D. & Delcourt, S. G. Thermodynamic effects of formamide on DNA stability. *Nucleic Acids Res.* **24**, 2095–2103 (1996).

98. Auer, A., Strauss, M. T., Schlichthaerle, T. & Jungmann, R. Fast, Background-Free DNA-PAINT Imaging Using FRET-Based Probes. *Nano Lett.* **17**, 6428–6434 (2017).

99. Abelson, J. *et al.* Conformational dynamics of single pre–mRNA molecules during in vitro splicing. *Nat. Struct. Mol. Biol.* **17**, 504–512 (2010).

100. Michelotti, N., de Silva, C., Johnson-Buck, A. E., Manzo, A. J. & Walter, N. G. Chapter Six - A Bird's Eye View: Tracking Slow Nanometer-Scale Movements of Single Molecular Nano-assemblies. in *Methods in Enzymology* (ed. Walter, N. G.) **475**, 121–148 (Academic Press, 2010).

101. Aitken, C. E., Marshall, R. A. & Puglisi, J. D. An Oxygen Scavenging System for Improvement of Dye Stability in Single-Molecule Fluorescence Experiments. *Biophys. J.* **94**, 1826–1835 (2008).

102. QUB - Markov Analysis. Available at: https://qub.mandelics.com/. (Accessed: 16th May 2018)

103. Armbruster, D. A. & Pry, T. Limit of Blank, Limit of Detection and Limit of Quantitation. *Clin. Biochem. Rev.* **29**, S49–S52 (2008).

104. Ultra-specific and amplification-free quantification of mutant DNA by single-molecule kinetic fingerprinting. submitted

105. Peterson, E. M. & Harris, J. M. Identification of Individual Immobilized DNA Molecules by Their Hybridization Kinetics Using Single-Molecule Fluorescence Imaging. *Anal. Chem.* **90**, 5007–5014 (2018).

106. Jones, P. A. & Takai, D. The Role of DNA Methylation in Mammalian Epigenetics. *Science* **293**, 1068–1070 (2001).

107. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).

108. Fraga, M. F. *et al.* The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. *Nucleic Acids Res.* **31**, 1765–1774 (2003).

109. Nogueira da Costa, A. & Herceg, Z. Detection of cancer-specific epigenomic changes in biofluids: Powerful tools in biomarker discovery and application. *Mol. Oncol.* **6**, 704–715 (2012).

110. Cheuk, I. W. Y., Shin, V. Y. & Kwong, A. Detection of Methylated Circulating DNA as Noninvasive Biomarkers for Breast Cancer Diagnosis. *J. Breast Cancer* **20**, 12–19 (2017).

111. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3**, 793–795 (2006).

112. Strauss, M. T., Schueder, F., Haas, D., Nickels, P. C. & Jungmann, R. Quantifying absolute addressability in DNA origami with molecular resolution. *Nat. Commun.* **9**, (2018).

113. Jungmann, R. *et al.* Multiplexed 3D Cellular Super-Resolution Imaging with DNA-PAINT and Exchange-PAINT. *Nat. Methods* **11**, 313–318 (2014).

114. McKinney, S. A., Joo, C. & Ha, T. Analysis of Single-Molecule FRET Trajectories Using Hidden Markov Modeling. *Biophys. J.* **91**, 1941–1951 (2006).

115. Blanco, M. R. *et al.* Single Molecule Cluster Analysis dissects splicing pathway conformational dynamics. *Nat. Methods* **12**, 1077–1084 (2015).

116. Christiansen, E. M. *et al.* In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images. *Cell* **173**, 792-803.e19 (2018).

117. Ouyang, W., Aristov, A., Lelek, M., Hao, X. & Zimmer, C. Deep learning massively accelerates super-resolution localization microscopy. *Nat. Biotechnol.* **36**, 460–468 (2018).

118. Teng, H. *et al.* Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience* **7**, (2018).

119. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).

120. Wei, B., Dai, M. & Yin, P. Complex shapes self-assembled from single-stranded DNA tiles. *Nature* **485**, 623–626 (2012).

121. Thubagere, A. J. *et al.* A cargo-sorting DNA robot. *Science* **357**, eaan6558 (2017).

122. Venkataraman, S., Dirks, R. M., Rothemund, P. W. K., Winfree, E. & Pierce, N. A. An autonomous polymerization motor powered by DNA hybridization. *Nat. Nanotechnol.* **2**, 490–494 (2007).

123. Dhakal, S. *et al.* Rational design of DNA-actuated enzyme nanoreactors guided by single molecule analysis. *Nanoscale* **8**, 3125–3137 (2016).

124. Multi-enzyme complexes on DNA scaffolds capable of substrate channelling with an artificial swinging arm | Request PDF. Available at: https://www.researchgate.net/publication/262610553_Multi-enzyme_complexes_on_DNA_scaffolds_capable_of_substrate_channelling_with_an_artificial_swinging_arm. (Accessed: 23rd January 2019)

125. A bio-hybrid DNA rotor–stator nanoengine that moves along predefined tracks | Nature Nanotechnology. Available at: https://www.nature.com/articles/s41565-018-0109-z. (Accessed: 21st January 2019)

126. Driessen, R. P. C. *et al.* Effect of Temperature on the Intrinsic Flexibility of DNA and Its Interaction with Architectural Proteins. *Biochemistry* **53**, 6430 (2014).

127. Kosaganov, Y. N. *et al.* Effect of Temperature and Ionic Strength on the Dissociation Kinetics and Lifetime of PNA−DNA Triplexes. *Biochemistry* **39**, 11742–11747 (2000).

128. Barakat, K., Issack, B. B., Stepanova, M. & Tuszynski, J. Effects of Temperature on the p53-DNA Binding Interactions and Their Dynamical Behavior: Comparing the Wild Type to the R248Q Mutant. *PLOS ONE* **6**, e27651 (2011).

129. Jang, H. S., Shin, W. J., Lee, J. E. & Do, J. T. CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function. *Genes* **8**, (2017).

130. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).

131. Kurdyukov, S. & Bullock, M. DNA Methylation Analysis: Choosing the Right Method. *Biology* **5**, (2016).