Programming temporal DNA barcodes for single-molecule fingerprinting

Shalin Shah,[†] Abhishek K Dubey,[‡,¶] and John Reif[*,†,‡]

†*Department of Electrical & Computer Engineering, Duke University, NC, US - 27701*

‡*Department of Computer Science, Duke University, NC, US - 27701*

¶*Computational Sciences and Engineering Division, Health Data Sciences Institute, Oak Ridge National Lab, TN, US - 37831*

E-mail: reif@cs.duke.edu

Abstract

Fluorescence microscopy enables simultaneous observation of the dynamics of single molecules in a large region of interest. Most traditional techniques employ either the geometry or the color of single molecules to uniquely identify (or barcode) different species of interest. However, these techniques require complex sample preparation and multicolor hardware setup. In this work, we introduce a time-based amplification-free single-molecule barcoding technique using easy-to-design nucleic acid strands. A dye-labeled complementary reporter strand transiently binds to the programmed nucleic acid strands to emit temporal intensity signals. We program the DNA strands to emit uniquely identifiable temporal signals for molecular-scale fingerprinting. Since the reporters bind transiently to DNA devices, our method offers relative immunity to photo-bleaching. We use a single universal reporter strand for all DNA devices making our design extremely cost-effective. We show DNA strands can be programmed for generating a multitude of uniquely identifiable molecular barcodes. Our technique can be easily incorporated with the existing orthogonal methods that use wavelength

or geometry to generate a large pool of distinguishable molecular barcodes thereby enhancing the overall multiplexing capabilities of single-molecule imaging.

Keywords

DNA kinetics, Nanoscale fingerprinting, Temporal patterns, DNA barcodes, Single-molecule imaging, TIRF

Optical multiplexing is defined as the ability to study, detect, or quantify multiple objects of interest simultaneously.[1–3] It has wide-ranging applications in fields such as data storage,[4,5] security and cryptography,[6,7] cellular and molecular-scale imaging,[1,2,8] high content multiplexed bio-assays,[9] color-display technologies,[10] detection of pathogens,[11] cytometrics,[12] and Lidar applications.[13] There are several ways to improve optical multiplexing, namely, using orthogonal wavelengths, multiple mesoscale geometries, orthogonal nucleic acid probes or a combination of these. The simplest technique is the use of orthogonal wavelengths and it is widely used for bio-imaging applications.[14–17] However, such techniques are limited by the number of spectrally distinct fluorescent colors and requires multicolor hardware setup. Contrary to simple wavelength multiplexing, by leveraging the geometry of a nanostructure and multiple fluorescent colors, Lin et al. self-assembled a large pool of fluorescent barcodes.[1] Similar to other geometry-based techniques, their barcode pool scales exponentially, however, the barcodes are several hundreds of micrometers tall and require complex nanostructure self-assembly.

A new class of passive imaging technique was introduced by the advent of super-resolution imaging technique, namely, the Exchange-PAINT.[2,18] They were able to achieve high single-color optical multiplexing by moving the reporter programming from dye color to the DNA sequence. Although their goal was high resolution fluorescence imaging, it can potentially be used for molecular barcoding. In order to perform single-color imaging, they imaged sequentially in time batches. Different reporter was used in each time batch to report the

regions containing the complementary DNA strands. A given reporter is washed off after imaging the assigned specie to be replaced by the next reporter. This process is repeated several times until all the species have been reported to reconstruct a high-quality image. This technique has a theoretical upper bound of 4^N on the number of available orthogonal imaging channels, where N is the number of nucleotide bases used in the fluorescent probe. Exchange-PAINT is a well-established technique, however, it requires multiple fluorescent reporters, sophisticated flow-chamber, and advanced drift-correction steps.

In this letter, we report a novel time-based fluorescence microscopy technique using simple nucleic acid devices for barcoding single-molecules. The basic workflow of our technique is illustrated in Fig. 1. We design short DNA strands, referred hereby as *DNA devices*, and attach them on a glass surface, as shown in Fig. 1a. The free-floating fluorescent reporters in the solution transiently bind to these designed devices and emit fluorescence upon binding, as shown in Fig. 1d. We design a set of DNA devices such that the hybridization kinetics produce distinct output temporal intensity signals, referred as *temporal DNA barcodes*. Although the observed temporal barcodes are stochastic, the underlining probability distribution is unique and can be identified from the temporal barcode, as shown in Fig. 1e. Our method is inspired from super-resolution imaging technique DNA-PAINT,[2] which also uses DNA strands and fluorescent reporters. However, unlike Exchange-PAINT, which achieves multiplexed single-color imaging by changing the sequence of fluorescent reporter, we keep the DNA sequences same and change device parameters such as the length and number of domains. This simple design decision makes data acquisition simpler, reduces the experimental cost and, yet, only requires single fluorescent channel. Such technique is compatible with prior techniques and can be easily incorporated into other works.[3,19,20] Note that in the DNA-PAINT literature, they refer to the short DNA strands on the nanostructure surface as the docker strands since their goal is to use these short strands as a docking platform for attaching reporters to do super-resolution imaging. Here, we refer to short strands attached on the glass surface as DNA devices since they are programmed to be used as taggants for

single molecule fingerprinting.

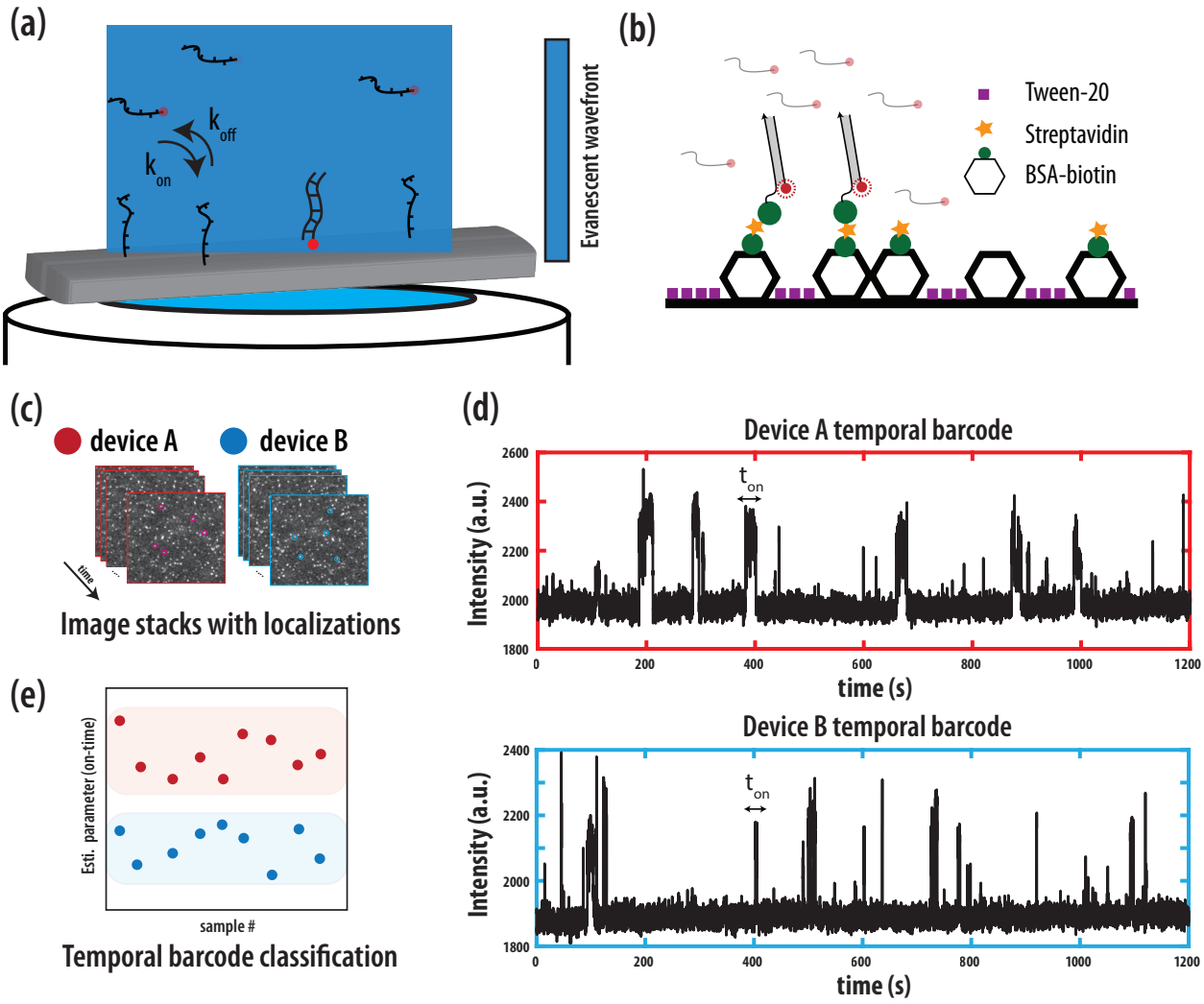

Figure 1: Overview of the temporal barcoding framework. (a) Several short DNA devices are attached on the glass surface. A complementary fluorescent reporter floats in the solution and reports the device activity transiently when it attaches to the device. TIRF microscopy is used for imaging as the evanescent wave-front can improve signal-to-noise ratio. (b) The devices are biotin-labeled and use biotin-streptavidin chemistry to attach on the glass surface. (c) Image stack are acquired for each device and localization coordinates are extracted to generate their intensity time trace shown in (d). (d) An illustration of a raw time signature for device A (10 nt) and device B (9 nt) collected at 10 Hz. (e) These temporal traces are analyzed in the parameter space plot to demonstrate their distinct behavior.

In fact, temporal signals have been used previously in literature for different studies. The upconversion nanocrystals by Lu et al.[7] and the chromophore network arrangement by Wang et al.[13] use the photon count signals for improved multiplexing. These studies, however, use

a more advanced detection technology with the capability of collecting photons at a much faster rate. Other works that demonstrate tunable time-signals include the single-molecule DNA-based clock and micro-RNA fingerprinting based on hybridization kinetics.[20,21] Several other kinetic studies have been conducted in the single-molecule FRET community using two dyes, one acting as a donor and the other as a receptor, with energy transfer between the dyes.[22,23] Although FRET probes offer much higher signal-to-noise ratio than our probing technique, the experimental design and acquisition hardware is much more complex because multiple detectors and a sophisticated dye choice are required.

For experimental verification, we design three simple ssDNA devices of length 8, 9 and 10 nt exploiting the well-established behavior of the DNA devices that their melting temperature scales with their length.[24,25] We modified these devices with a biotin-label on one end to attach them on a glass surface using a streptavidin molecule attached to BSA-biotin. Imaging was performed using an inverted fluorescence microscope in total internal reflection fluorescence (TIRF) mode. This offers the benefit of relatively higher signal-to-noise ratio as the background scattering due to the free-floating fluorescent reporters is minimized. The free-floating fluorescent strands were labeled with an ATTO 647N dye (646 nm excitation peak). In order to extract the temporal barcodes from the raw image stack, a multistage pipeline is introduced, as shown in supplementary Fig. S5. This pipeline detects the localization coordinates, extracts the temporal barcodes and denoises using machine learning. For full experimental details such as sample preparation, data acquisition, nucleotide sequence of devices, and image processing pipeline, refer to the supplementary methods and tables.

A simple ssDNA device will be temporarily fluorescent when the reporter attaches, and it will go dark once the reporter dehybridizes. If the activity of this device is observed over time, we should observe a noisy intensity trace as shown Fig. 2b. This process can be modeled with a two state Markov chain, as demonstrated in theoretical studies.[26] A compact reaction network of the process is shown in Fig. 2a along with how device looks like in each state. An denoised signal with two states, constructed using the mean shift clustering al-

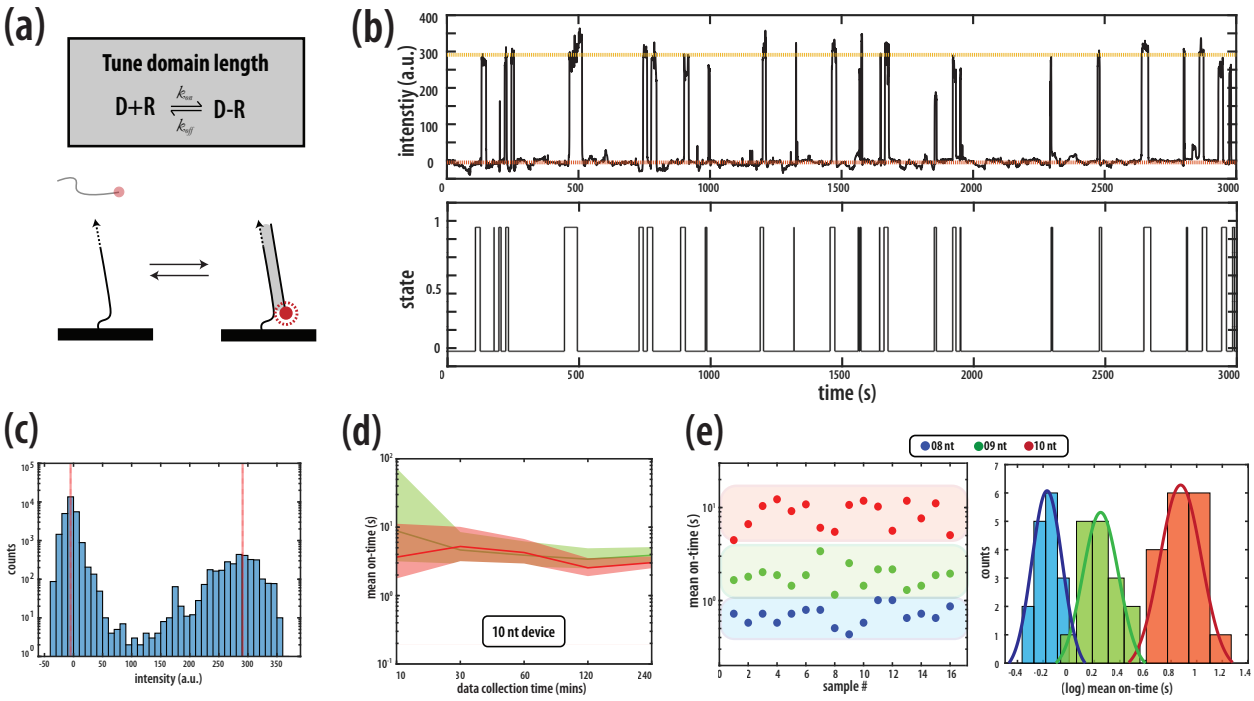

Figure 2: Programming the length of DNA device to tune their temporal DNA barcodes. (a) The chemical reaction showing the binding and unbinding of the fluorescent reporter with a short DNA device. The dotted lines demonstrate modifying the length to program device's temporal DNA barcode. (b) A sample temporal DNA barcode collected from a device with length 10 nt. Mean shift clustering is applied on the filtered signal to find two peaks (on-state and off-state) and obtain the hidden state chain. (c) The histogram of the temporal barcode shown in (b) along with the identified peak locations. (d) A control experiment demonstrating a tighter bound on the estimated on-time of a single 10 nt device with increased data collection time (or the number of collected samples). The same experiment was conducted twice to verify reproducible behavior. (e) Data for three devices, namely, 8 nt, 9 nt and 10 nt was collected and the estimated on-time for multiple localizations was analyzed. Full scatter plot and histogram counts are shown for visualization purposes demonstrating their distinguishable behavior.

gorithm,[27] is shown in Fig. 2b. The dotted lines indicates the location of peaks identified by the unsupervised clustering algorithm. The histogram of intensity counts along with the identified peak location is also shown in Fig. 2c. We then estimate the on-times for each blink from denoised barcode for device classification (refer to supplementary Fig. S6 for the definition of on-time in a temporal DNA barcode).

First, we identify an operable device and reporter concentration range. In particular, we choose low device concentration to avoid the spatial overlap of multiple devices in a diffrac-

tion limited region. The reporter, however, was kept much higher concentration than the device to allow for the continuous replenishment of reporters from the solution. The laser intensity was set sufficiently low to ensure relative immunity to photo-bleaching (refer supplementary Fig. S3).

We next acquire microscopy data for 8 nt, 9 nt and 10 nt devices to verify the distinguishability of devices, previously predicted by the theoretical studies.[26] We analyze several localization spots for each device using our software scripts and plot the estimated mean on-time for each device at several data collection times. We observe that the estimated mean on-time for several localizations of each device has higher variance for shorter data collection time (refer to supplementary Fig. S8). However, as the data collection time is increased to 60 minutes, a better separation is achieved, as shown in Fig 2e.

We also studied the variance of on-time estimates for the device population. We observe that the variance of the on-time estimate decreases as expected, as we collect longer temporal barcodes. The rate of decrease of the variance was fast enough to achieve the separation in the devices of different length by simple state-of-the-art clustering methods such as k-means. In other words, we can use these device as a single molecule barcodes for molecular-scale fingerprinting. The expected exemplary temporal barcode for all three devices is shown in supplementary Fig. S9 for reference purposes.

Next, we analyze some exemplary temporal barcodes of a device to verify whether the on-time random variable follows the exponential distribution, as modeled by the Markov chain in prior theoretical works.[26] Once verified (refer to supplementary Fig. S7), we used the model to estimate the 95% confidence interval of our on-time estimates. In Fig. 2d, we observe that the 95% confidence interval of the mean value estimate gets tighter and tighter as the data collection time increased. This implied that the on-time parameter can be reliably used for the task.

Some additional remarks on the on-time estimates for the simple devices are in order. First, we observed an up-shift in the estimates than our theoretical estimates.[26] However, the the-

oretical study ignored the hardware limitations in their calculations. In particular, they ignored the loss of short lived events because of the detector's limit on the rate of data acquisition especially at high signal-to-noise ratio. Nevertheless, the rough exponential dependence of DNA hybridization kinetics on its length generates distinct temporal patterns, as observed in prior sections, demonstrating the robustness of our framework.[20,22,24]

After successfully classifying DNA devices using the domain length, we extended our barcoding idea on multiple domains. We designed new DNA devices each containing two domains where the length of each domain can be tuned to program the output temporal DNA barcode. An additional domain means up to two reporters can attach to one device at the same time, leading to an additional bright state in the temporal barcode. The dynamic flow of a double-domain device along with the simplified reaction network is shown in Fig. 3a. As shown in the figure, the fluorescent reporter can attach to either domain initially leading to a fluorescent state. If another reporter also attaches to the device during this time, the fluorescent intensity will nearly double[28] as both the reporters are attached at the same time. Note that we cannot distinguish which domain the reporter attaches to, and instead we only observe two fluorescent states.

We designed three devices with double-domain: 9-9 nt, 9-10 nt and 10-10 nt. Each sample was prepared for imaging, however, the device concentration was further reduced to account for the increased fluorescence activity due to double domain. The expected typical temporal intensity signals for all three devices is shown in supplementary Fig. S10. The collected data was processed to extract temporal barcode using our information extraction pipeline. The denoised three-state approximate state chain was analyzed for two parameters: on-time and double-blink time (refer supplementary Fig. S6 for definition). The results of the estimated mean on-time and double-blink time for several localizations of each device are shown in Fig. 3b.

The temporal signatures of devices 9-9 nt and 9-10 nt are separated by their overall fluorescent-ON activity whereas the temporal signatures of devices 9-10 nt and 10-10 nt are separated

by their double-blink activity. Note the use of two parameters to analyze a temporal intensity signal helps with the faster separation of DNA devices even if the data collection time is short. Clearly, from the Fig. 3b at our current data collection time, a single parameter on-time cannot distinguish the 9-10 nt and 10-10 nt device, which is not the case when we analyze it in the second dimension. Such additional parameter extraction can be especially useful when the temporal signal is multi-dimensional or longer data collection is difficult. Finally, the variance of the estimated data points can be easily reduced by longer data collection times (refer supplementary Fig. S8), faster detection hardware and stronger laser powers.[25,29]

Several other devices can be easily designed using our idea of tuning the number of domains and length as seen in the prior theoretical work.[26] The desired set of devices depend on the available hardware setup and detection technology. If the hardware has fast detection capability and high power laser, shorter devices are preferred. If the hardware has slower detection capability and low power laser, longer devices are preferred as the overall process becomes slower. This makes our time-based framework easy to adopt and versatile. If further tuning is desired, more domains can be added to the device, however, the task of classifying the collected temporal barcodes will require training a machine learning model as the decoding process becomes more complex.[26]

In addition to programming the bright-time activity, we controlled the dark time of devices by using a secondary structure such as DNA hairpin. Prior works have tuned the dark time of temporal intensity traces by controlling the reporter concentration as the binding process follows a second order kinetics.[24] In this work, we instead exploit the programmable secondary structure of DNA to achieve this. It is well-known that if a DNA sequence contains a complementary sub-sequence then it can self-fold into a hairpin-like structure.[30–32] As shown in Fig. 3c, we change the neck length of the hairpin to control the availability of the domain complementary to the fluorescent reporter. This is because the reporter can only bind to the hairpin neck when it is open thereby reducing the total available device time (or increasing

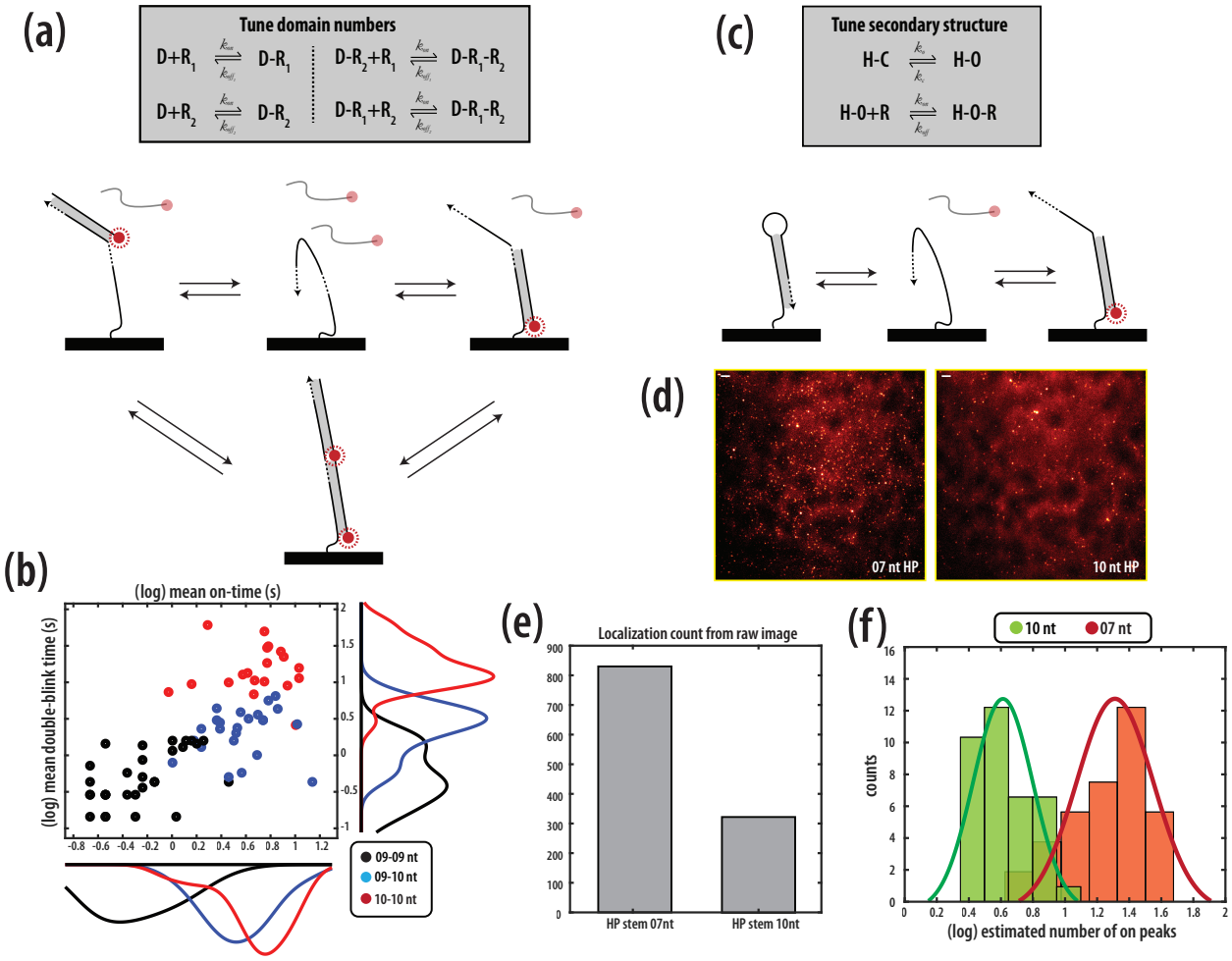

Figure 3: Programming the secondary structure of DNA devices and the number of reporter domains. (a) The chemical reactions showing different states for tuning the number of domains along with their lengths. (b) Several localizations of three devices were analyzed for their on-time and double-blink time to verify their distinguishable behavior. (c) The chemical reactions showing different states for tuning the length of a secondary structure. (d - e) The raw unprocessed images of each device along with their spot count indicating that longer hairpin lengths reduce the probability of reporter binding (or increases the hairpin downtime). (f) A histogram plot containing the analyzed number of on-peaks for several localizations of each device. Scale bars in (d) are 5 µm.

the device downtime).

To demonstrate this idea, we designed two hairpin devices with neck length of 7 nt and 10 nt (for sequence details, refer to supplementary tables). The device with a neck length of 7 nt should behave almost like a standard ssDNA device as the average binding time of such DNA device is very short.[18,22] The device with neck length of 10-nt should have large chunks

of down times as DNA devices with 10 nt hybridize for a much longer time on average. As expected, the difference in the number of fluorescent localizations is directly visible from the raw image data of both the devices shown in Fig. 3d and Fig. 3e. The number of localizations in a typical frame of 7 nt device is much higher than 10 nt device at a given device and reporter concentration demonstrating the increased down time of 10 nt device. Additionally, after processing the image stack of these devices using our information extraction pipeline, there is also a clear difference observed in the number of blinking peaks observed for the respective temporal barcodes as shown in Fig. 3f. The number of fluorescent-ON peaks in a given time is much less for a device with longer stem length, as seen in the Fig. 3f, acting as a parameter for easy classification of these devices.

Although our temporal framework offers a new potential venue to study single molecule, there are several important challenges in the field that should be noted. Longer data acquisition, and higher signal-to-noise ratio is always desirable and while this is possible to some extent by tuning the laser power, detector gain, capture rate and better dyes, there is an upper wall. A potential solution to analyze such under-sampled and noisy data includes training a machine learning model such as generative adversarial networks (GAN)[17,27] on the microscopy data for image denoising and signal identification process. This has a potential to further improve our software pipeline process by making it faster and robust. Additionally, by learning the structure of data from the noisy stack, it may be possible to be able to reconstruct high-quality signals even if data is highly under-sampled. The field of super-resolution imaging has already seen benefits of deep learning[17] and these can applied to a broader field of DNA nanoscience and single-molecule imaging.

In conclusion, we have introduced a new reporting framework that uses the time domain for barcoding single molecules. Our framework consists of a programmable simple DNA device that can be attached to the specie of interest and a complementary universal reporter strand that can transiently bind with DNA devices to report their activity. The device reporter combination together creates a temporal barcode as the temporal intensity signals emitted

are distinct and can act as a fingerprint. We move the programming from fluorescent reporter and dye color to the DNA devices as our goal is to develop an economical and simple-yet-powerful method. Our framework can work with as few as one dye, greatly simplifying the hardware setup and data acquisition for single molecule systems. The introduction of the use of the time domain for reporting provides an additional channel for multiplexing that can substantially increase the number of barcodes. Further, prior techniques such as use the geometry of a nanostructure or multiple dyes, can be easily incorporated with our work to scale the number of molecular barcodes exponentially. Finally, we have also designed a set of open-source MATLAB script to simplify the data extraction from image stack which we believe will aid beyond our barcoding framework to the general community of fluorescence microscopy and super-resolution imaging.

In general, the ability to design a reporting system that can report real-time breathing activity of a cellular or molecular object can offer numerous future applications in the nanoscience field. For example, this can be used to observe real-time behavior of localized DNA circuits by carefully integrating our reporting DNA devices with the existing logic circuits.[30] Several other applications such as error-correcting temporal barcodes and cellular tagging are also possible.[26]

Author contributions

S.S conducted the study, performed experiments and wrote the paper. A.D analyzed the results and wrote the paper. J.R initiated the study and supervised it.

Acknowledgement

The authors thank Yasheng Gao of Duke LMCF for imaging advice with TIRF scopes. The authors would like to thank Daniel Fu, Xin Song, Ming Yang, Tianqi Song, Abeer Eshra and Hieu Bui for their comments and suggestions. The authors would also like to thank Xin

Song and Isaiah Wales for proof-reading the manuscript.

This work was supported by National Science Foundation Grants CCF-1813805 and CCF-1617791.

Supporting Information Available

Detailed description of materials and methods and additional figures and tables. This material is available free of charge via the Internet at `http://pubs.acs.org`.

The scripts developed to work with microscopy data are also available as part of the supplementary files. The most up-to-date files and status of the scripts can be found online on the GitHub repository: `https://github.com/shalinshah1993/temporalDNAbarcodes`

References

(1) Lin, C.; Jungmann, R.; Leifer, A. M.; Li, C.; Levner, D.; Church, G. M.; Shih, W. M.; Yin, P. *Nat. Chem.* **2012**, *4*, 832.

(2) Jungmann, R.; Avendaño, M. S.; Woehrstein, J. B.; Dai, M.; Shih, W. M.; Yin, P. *Nat. Methods* **2014**, *11*, 313.

(3) Jungmann, R.; Avendaño, M. S.; Dai, M.; Woehrstein, J. B.; Agasti, S. S.; Feiger, Z.; Rodal, A.; Yin, P. *Nat. Methods* **2016**, *13*, 439.

(4) Mottaghi, M. D.; Dwyer, C. *Adv. Mater.* **2013**, *25*, 3593–3598.

(5) Zijlstra, P.; Chon, J. W.; Gu, M. *Nature* **2009**, *459*, 410.

(6) Nellore, V.; Xi, S.; Dwyer, C. *ACS nano* **2015**, *9*, 11840–11848.

(7) Lu, Y. et al. *Nat. Photonics* **2014**, *8*, 32.

(8) Hu, F.; Zeng, C.; Long, R.; Miao, Y.; Wei, L.; Xu, Q.; Min, W. *Nat. Methods* **2018**, *15*, 194.

(9) Nguyen, H. Q.; Baxter, B. C.; Brower, K.; Diaz-Botia, C. A.; DeRisi, J. L.; Fordyce, P. M.; Thorn, K. S. *Adv. Opt. Mater.* **2017**, *5*, 1600548.

(10) Deng, R.; Qin, F.; Chen, R.; Huang, W.; Hong, M.; Liu, X. *Nat. Nanotechnol.* **2015**, *10*, 237.

(11) Li, Y.; Cu, Y. T. H.; Luo, D. *Nat. Biotechnol.* **2005**, *23*, 885.

(12) Krutzik, P. O.; Nolan, G. P. *Nat. Methods* **2006**, *3*, 361.

(13) Wang, S.; Vyas, R.; Dwyer, C. *Opt. Express* **2016**, *24*, 15528–15545.

(14) Sharonov, A.; Hochstrasser, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18911–18916.

(15) Huang, B.; Wang, W.; Bates, M.; Zhuang, X. *Science* **2008**, *319*, 810–813.

(16) Schnitzbauer, J.; Strauss, M. T.; Schlichthaerle, T.; Schueder, F.; Jungmann, R. *Nat. Protoc.* **2017**, *12*, 1198.

(17) Ouyang, W.; Aristov, A.; Lelek, M.; Hao, X.; Zimmer, C. *Nat. Biotechnol.* **2018**,

(18) Woehrstein, J. B.; Strauss, M. T.; Ong, L. L.; Wei, B.; Zhang, D. Y.; Jungmann, R.; Yin, P. *Sci. Adv.* **2017**, *3*, e1602128.

(19) Werbin, J. L.; Avendaño, M. S.; Becker, V.; Jungmann, R.; Yin, P.; Danuser, G.; Sorger, P. K. *Sci. Rep.* **2017**, *7*, 12150.

(20) Johnson-Buck, A.; Su, X.; Giraldez, M. D.; Zhao, M.; Tewari, M.; Walter, N. G. *Nat. Biotechnol.* **2015**, *33*, 730.

(21) Johnson-Buck, A.; Shih, W. M. *Nano Lett.* **2017**, *17*, 7940–7944.

(22) Tsukanov, R.; Tomov, T. E.; Masoud, R.; Drory, H.; Plavner, N.; Liber, M.; Nir, E. *J. Phys. Chem. B* **2013**, *117*, 11932–11942.

(23) McKinney, S. A.; Joo, C.; Ha, T. *Biophys. J.* **2006**, *91*, 1941–1951.

(24) Jungmann, R.; Steinhauer, C.; Scheible, M.; Kuzyk, A.; Tinnefeld, P.; Simmel, F. C. *Nano Lett.* **2010**, *10*, 4756–4761.

(25) Mucksch, J.; Blumhardt, P.; Strauss, M. T.; Petrov, E. P.; Jungmann, R.; Schwille, P. *Nano Lett.* **2018**, *18*, 3185–3192.

(26) Shah, S.; Reif, J. Temporal DNA Barcodes: A Time-Based Approach for Single-Molecule Imaging. International Conference on DNA Computing and Molecular Programming. 2018; pp 71–86.

(27) Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep learning*; MIT press Cambridge, 2016; Vol. 1.

(28) Schmied, J. J.; Raab, M.; Forthmann, C.; Pibiri, E.; Wünsch, B.; Dammeyer, T.; Tinnefeld, P. *Nat. Protoc.* **2014**, *9*, 1367.

(29) Lee, T.-H. *J. Phys. Chem. B* **2009**, *113*, 11535–11542.

(30) Bui, H.; Shah, S.; Mokhtar, R.; Song, T.; Garg, S.; Reif, J. *ACS nano* **2018**, *12*, 1146–1155.

(31) Eshra, A.; Shah, S.; Song, T.; Reif, J. *IEEE Trans. Nanotechnol.* **2019**, *18*, 252–259.

(32) Garg, S.; Shah, S.; Bui, H.; Song, T.; Mokhtar, R.; Reif, J. *Small* **2018**, *14*, 1801470.