Supporting information: 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

Supplementary Methods

Materials

All the DNA strands - biotin-labeled, dye-labeled and unlabeled - were ordered from IDT DNA Technologies, USA. No PAGE purification was performed on the biotin-labeled strands. DNA strands labeled with fluorophores were ordered with HPLC purification. Biotin-BSA (Catalog no.29130-25mg) was purchased from Thermo-Fisher Scientific. Protocatechuic acid (PCA, Catalog no.37580-25G-F), Protocatechuate-3,4-dioxygenase (PCD, Catalog no. P8279-25UN), Streptavidin (Catalog no. S4762-5MG) was purchased from Sigma-Aldrich (St. Louis, MO). Color-frost glass-slides (Catalog no.10118-958) and cover-slips #1.5 (Catalog no.16002-256) were purchased from VWR International (Radnor, PA). Gold nanoparticles (Catalog no.753688-25ML) used for drift-estimation were purchased from Sigma-Aldrich. Immobilization buffer (10 mM Tris-HCl, 100 mM NaCl, 0.05% Tween-20, pH 7.5), imaging buffer (5 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM EDTA, 0.05% Tween-20, pH 8) and super-imaging buffer (5 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM EDTA, 0.05% Tween-20, pH 8 with ROXS system containing 1 mM PCA, 1 mM PCD, and 1 mM TX) were used in this work.

Sample preparation for imaging

A custom-built flow-chamber was prepared every-time, before sample imaging under TIRF microscope. To prepare a flow-chamber, two double-sided tapes were attached on glass-slide roughly 1 cm apart to glue cover-slip to the glass-slide. The volume between cover-slip and glass-slide is roughly 30 µL. The excess tape was removed, and the flow-chamber was cleansed by flowing 40 µL immobilization buffer. For surface immobilization of DNA, 20 µL of 1 mg mL^{-1} biotin labeled Bovine Serum Albumin (BSA) was flown and incubated for 2 minutes. After 2 minutes, excess biotin-BSA was removed by rinsing with 40 µL immobilization buffer. 20 µL streptavidin was flown through the chamber and incubated for 2 minutes at 0.5 mg mL^{-1} concentration to attach with biotin. The flow-chamber was again rinsed with 40 µL immobilization buffer and then with 40 µL imaging buffer. After the rinse, 20 µL of biotin-labeled DNA solution was flown through the chamber and incubated for 5 minutes (exact concentration varied for each device, in general 20 - 150 pM). To remove unbound DNA, the flow-chamber was cleansed with 40 µL of imaging buffer and then 100 nm gold-nanoparticles were flown through the chamber and incubated for 5 minutes. The sample was washed with imaging buffer and, then, with super-imaging buffer. Finally, 20 µL of fluorescence strands at 10 nM were introduced in the flow-chamber and then the sample was sealed with nail polish. Note that the concentration of each reagent and incubation times can be tuned to achieve desired separation of localizations in the final field of view. A general rule to follow is to tune these such that several tens of localizations are observed in each frame without a spatial overlap.

Optical setup

All the images were collected using a Leica DMI 6000B motorized inverted fluorescence microscope in TIRF mode with an oil-immersed objective lens (100x, N.A. = 1.46). To record the fluorescence from the devices, a Hamamatsu (C9100-13) Electron Multiplying Charge Coupled Device (EM-CCD) was used because of its capability of collecting single photons. The cell-size for EM-CCD is 16 µm × 16 µm and it captures an image of 512 × 512 pixels. By using a 100x objective, we were able to achieve a pixel size of 224 nm. To excite dyes, the setup has 3 laser diodes with 488, 561, and 635 nm were available. Filtering at the input and output was done using VBG triple filer (Volume Bragg Gratings - x490/20 d435 m465/45 — x422/44 d505 m545/55 — x552/24 d550 m610/65) and Cy5 filter (x 620/60 d660 m700/75). In this work, we only work with an ATTO 647N dye so we use 635 nm laser channel and Cy5 filter. To avoid unnecessary photo-bleaching, as shown in supplementary Fig. S3, minimal source laser intensity was used (approx. 10 mW, max capacity 18 mW). Several penetration depths and TIRF angles were tried to achieve the best contrast and maximum illumination. The fluorescence setup used along with the ATTO 647N fluorescence curve can also be visualized in supplementary Fig. S1.

Mitigating nonspecific binding

An important challenge before accurate extraction of time-signals is nonspecific binding. Nonspecific binding refers to the fluorescence activity arising from everything else except attached devices. In our case, the free-floating fluorescent strands produce a high level of nonspecific binding. To address this, we use biotin-labeled BSA (1:10 biotin-labeled BSA: BSA) and Tween 20 (0.1 % (v/v) in buffers) as it is shown to reduces the amount of non-specific localizations over extended time.[1–3] Although BSA and Tween 20 greatly mitigates the nonspecific binding, it does not eliminate it completely. The fluorescence intensity trace of a typical nonspecific binding and a device is shown in supplementary Fig. S11 for reference purposes. Clearly, there is a difference in the number of peaks observed between the real signal and nonspecific signal. Therefore, we incorporate analyzing time-signatures of each localization in our information extraction pipeline (mean z-projection and peak filter step) to accurately discard several localization that can potentially represent nonspecific binding. Exchange-PAINT study also used a preliminary software-based nonspecific binding elimination.[4]

Cluster machine configuration

The cluster machine used to run the MATLAB scripts for barcode extraction and localization detection had the following configuration: 10x Tensor TXR231-1000R D126 Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz (512GB RAM - 40 cores). Note that the use of cluster machine is crucial since most raw data files are several hundreds of gigabytes making it extremely difficult to handle them. Therefore, in order to avoid dealing with memory overflow issue, we use machines with much larger available main memory.

Information extraction pipeline

There are several steps involved to extract meaningful information from the recorded raw image stacks. An abstract pipeline of this process is shown in supplementary Fig. S5a. As shown in the figure, the first step is data-collection *i.e* recording an image stack using TIRF microscope. Once we have the raw data, we convert the proprietary Leica lif file to a mat file using the bfmatlab library. This can help us with the development of the programmable downstream MATLAB scripts as the raw data is now available in the supported format.

Drift-estimation and correction

The next step includes the estimation and correction of the lateral (x, y-direction) and axial (z-direction) drift. For lateral drift correction, we use the redundant cross-correlation algorithm proposed by Wang et al.[5] by incorporating their library within our MATLAB script. In order to avoid major axial drift correction, we first waited for a few minutes for the sample to settle and also applied a software-based hardware focus assist available in the proprietary Leica suite. For further axial drift removal, we calculate the average intensity for each frame and fit a polynomial curve to the average intensity trace over time, as shown in supplementary Fig. S4b, as this can be the estimated baseline value for each frame. The estimated baseline is then subtracted from the image stack. The effect of this process on a sample temporal intensity trace is shown in supplementary Fig. S4c.

Extracting localization coordinates

Once the drift corrected data stack is available, we apply several filters to locate the localizations and find their centroid coordinates. The sub-pipeline shown in supplementary Fig S5b summarizes the filter steps involved in this process. Note that we call this process a fast localization detection step since we use only one frame to find all the localization locations. The standard process in the field of single-molecule localization microscopy is to find frame-wise localizations as their goal is to obtain all the coordinates and then super-impose them. In our case, since we only need the coordinate of a DNA device in any one of the frames, the process can be expedited by simply finding the localizations in the summation z-projection frame. This also helps discard some of the nonspecific spots as they will only be fluorescent in a few frames making their z-projection sum value small.

Extracting temporal intensity traces

After extracting the possible set of device coordinates, the temporal intensity time trace is generated assuming the point spread function of 3 × 3 pixels. The sub-pipeline in Fig.S5c shows several steps involved in the barcode extraction and analysis process. After obtaining the intensity time trace for each localization, the next step includes applying the wavelet filter. It was found that unlike other bandpass filters, wavelet filter performs exceptionally well in removing shot noise[6] as seen in supplementary Fig. S4c. The filtered temporal barcodes are clustered in two or three states depending on the device using the unsupervised mean shift clustering technique to obtain a state chain. Prior smFRET works have used the Hidden Markov model (HMM) to denoise time-series data,[7,8] however, we found its output rather subjective since our signal-to-noise ratio is much lower than FRET probes. Additionally, the uneven background illumination leads to a wide dynamic range of fluorescent intensity values for temporal barcodes. Therefore, we use simple and fast unsupervised mean shift clustering for approximating the denoised signal from the noisy temporal barcodes. This state chain can be analyzed to extract parameters such as dark-time, on-time, double-blink etc.

Unpurified DNA device

All the data collected in this work was from unpurified DNA devices. This is because in order to distinguish a small pool of DNA devices with an exponential difference in their hybridization kinetics makes the PAGE purification non-essential. Our longest device is 23 nt (including the poly-T padding), and the maximum pool-size of devices is 8. While no PAGE purification makes the data-collection process simpler, it introduces some error in the estimated on-times. The shortest and longest device used in this work had a synthesis efficiency of approximately 91% and 87% respectively assuming a coupling efficiency of 99.30%, as suggested by IDT tables.[9] In order to mitigate the effects of truncation, we employ careful design of DNA strands.

Current synthesis technology extends from $3'$ to $5'$ direction so the truncation is usually at $5'$-end.[3] Since we attach biotin labels on $5'$ end, a truncated DNA strand will not have a biotin label and, therefore, will simply flow out of imaging chamber during sample preparation. As a result, the overall percentage of imaging a truncated device is substantially low. Therefore, even without purification, we can still distinguish the temporal barcodes of our devices as seen in Fig. 2 and Fig. 3. However, when the device pool becomes much larger (for example, using DNA origami to design a pool of 50 devices), PAGE purification of individual devices will become essential as the variance in the estimated parameters (on-time) is required to be very low.

Supplementary tables

Table S1: A complete list of all the DNA sequences used in this work.

Name	**Sequence (5′ to 3′ direction)**
Dye-labeled reporter	CTAGATGTAT-/ATTO647N/
D1-ssDNA-10nt	/5BiosG/-TTTTTT-ATACATCTAG
D2-ssDNA-09nt	/5BiosG/-TTTTTT-ATACATCTA
D3-dd-10nt-10nt	/5BiosG/-TTTT-ATACATCTAG-T-ATACATCTAG
D4-dd-10nt-09nt	/5BiosG/-TT-ATACATCTAG-T-ATACATCTA
D5-dd-09nt-09nt	/5BiosG/-TT-ATACATCTA-T-ATACATCTA
D6-hp-10nt-10nt	/5BiosG/-TT-ATACATCTAGTTTTTTCTAGATGTAT
D7-hp-10nt-07nt	/5BiosG/-TT-ATACATCTAGTTTTTTTATGTAT

Table S2: Localization count after each step in the filter process

Name	**Z-projected image**	**Meanshift filtered**	**Peak filtered**	**Human filtered**
08nt	117	36	17	17
09nt	149	78	49	17
10nt	214	83	42	17

Supplementary figures

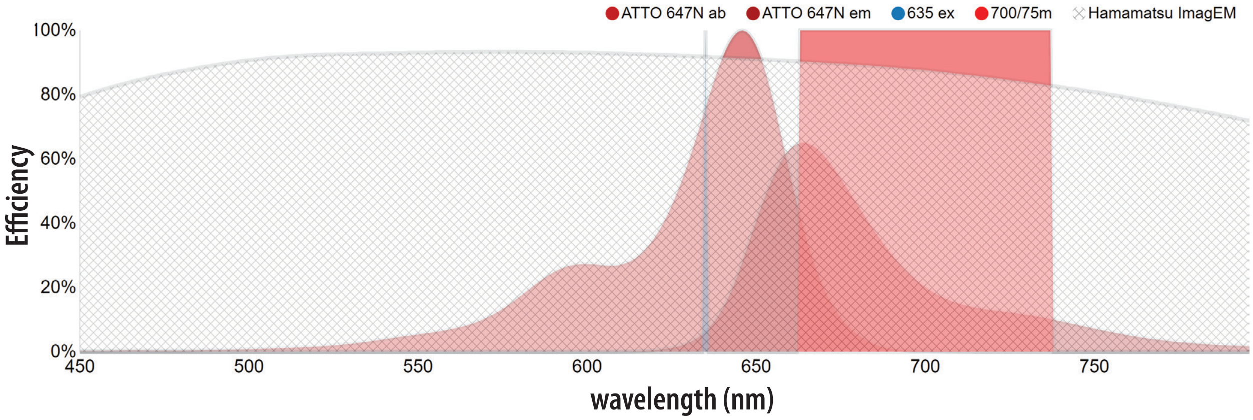

Figure S1: The fluorescence spectra viewer for the experimental configuration used for data-collection. It shows the absorbance and emission spectrum of ATTO 647N dye, along with the 635 nm laser line and emission band-pass filter. The grey line shows collection efficiency of Hamamatsu EM-CCD camera. The emission curve is scaled to account for the quantum yield. This curve is generated using FPbase Fluorescence Spectra Viewer.[10]

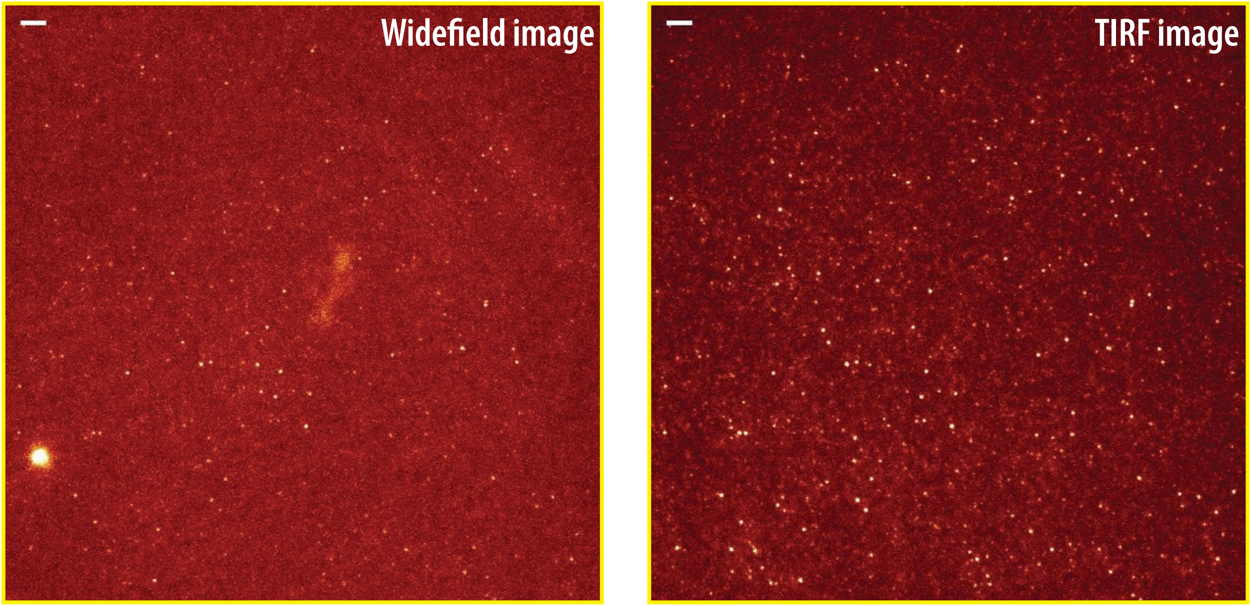

Figure S2: A sample widefield and TIRF image of the sample demonstrating increased signal-to-noise ratio due to the evanescent wavefront. All the scale bars are 5 µm.

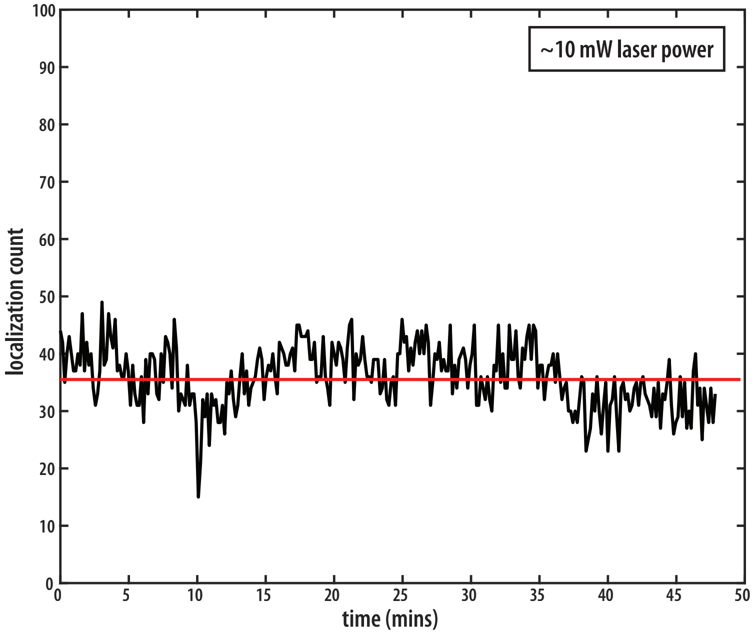

Figure S3: A localization count graph demonstrating the relative immunity to photobleaching. The 10 nt device was used to construct this plot as it is the longest domain length used in this work. Shorter device are less likely to bleach since their average binding times are shorter. Therefore, we used about 10 mW laser power at the source (maximum capacity 18 mW) for all the data-collection experiments. Such relative immunity to photobleaching is also observed by some prior quantitative super-resolution imaging techniques.[11] Note that this graph was generated after filtering each ROI image by applying a bandpass filter.

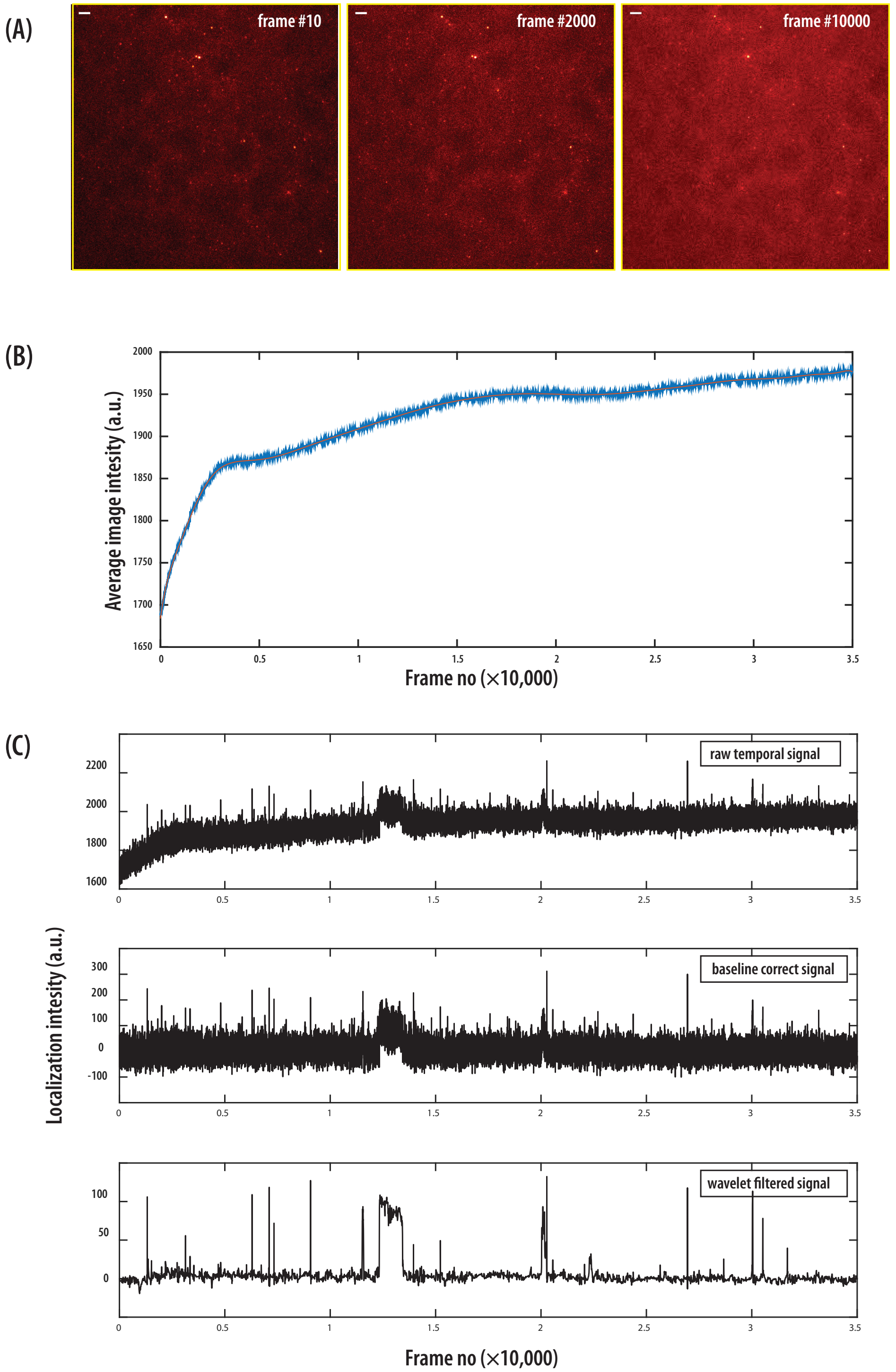

Figure S4: A sample data used to demonstrate the filter process. (a) For longer acquisition times, although there is software-based real-time z-drift correction, there is still a shit in the dynamic range of pixel values. (b) Average image intensity plot over time (or frame count) shows the z-drift. (c) Sample temporal barcodes for raw data, after baseline correction and after using wavelet filter. All the scale bars are 5 µm.

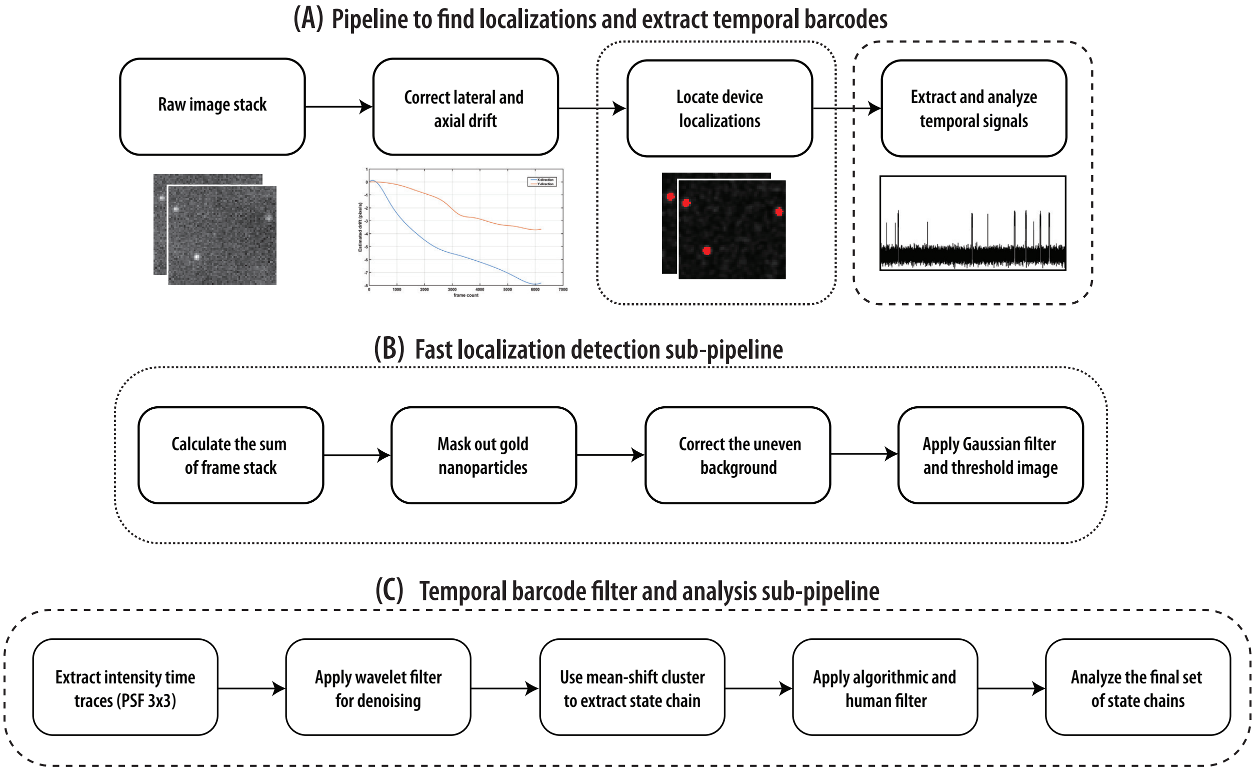
(A) Pipeline to find localizations and extract temporal barcodes
Raw image stack
Correct lateral and axial drift
Locate device localizations
Extract and analyze temporal signals
(B) Fast localization detection sub-pipeline
Calculate the sum of frame stack
Mask out gold nanoparticles
Correct the uneven background
Apply Gaussian filter and threshold image
(C) Temporal barcode filter and analysis sub-pipeline
Extract intensity time traces (PSF 3x3)
Apply wavelet filter for denoising
Use mean-shift cluster to extract state chain
Apply algorithmic and human filter
Analyze the final set of state chains

Figure S5: The pipeline to extract temporal barcodes, analyze them and extract useful parameters. (a) The global pipeline showing several steps performed on the raw image stack to extract temporal barcodes. This includes sub-steps such as the drift correction and the localization coordinate extraction. (b) The localization detection step has a sub-pipeline which includes steps such as removing gold nanoparticles and correcting uneven background illumination. (c) The barcode extraction step also has several steps shown in the sub-pipeline. This includes intensity time trace calculation from a 3×3 pixel area, application of denoising filters and a careful selection of the state chains for final analysis.

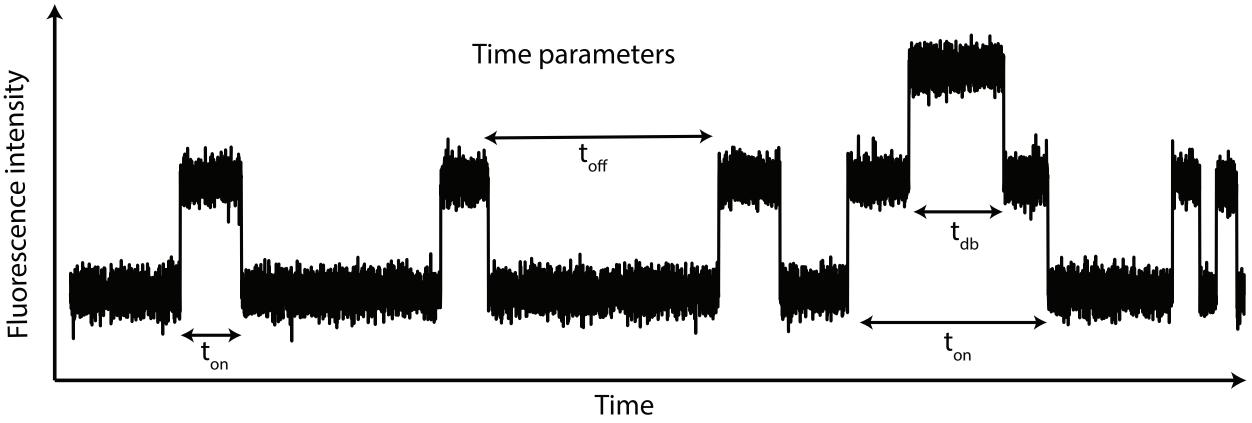

Figure S6: A sample simulated temporal intensity trace showing several parameters such as the on-time (or bright time), off-time (or dark time) and double-blink time. Simulations were conducted by using scripts from prior work on temporal DNA barcodes by Shah et al.[12]

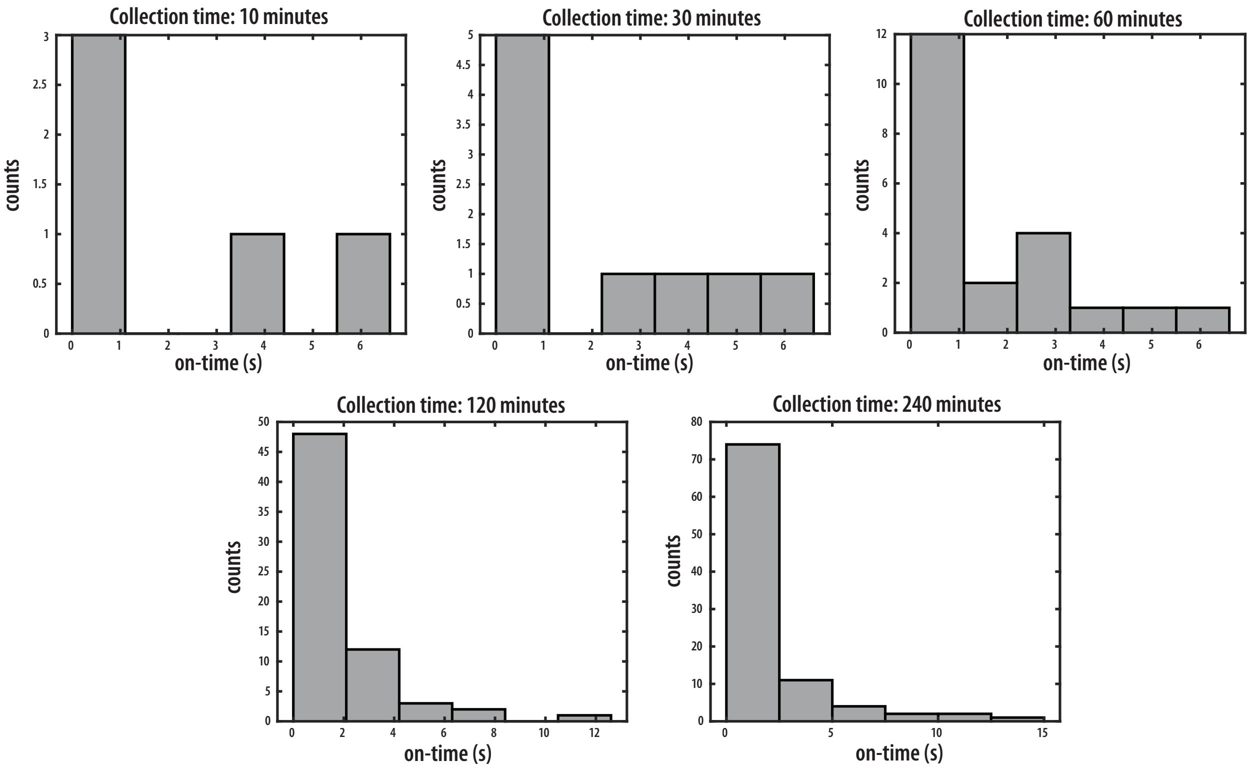

Figure S7: A histogram constructed from the calculated on-time values of a temporal barcode at different data collection times. Clearly, as the data collection time increase the histogram resembles more to an exponential distribution and can be fit to one with higher confidence.

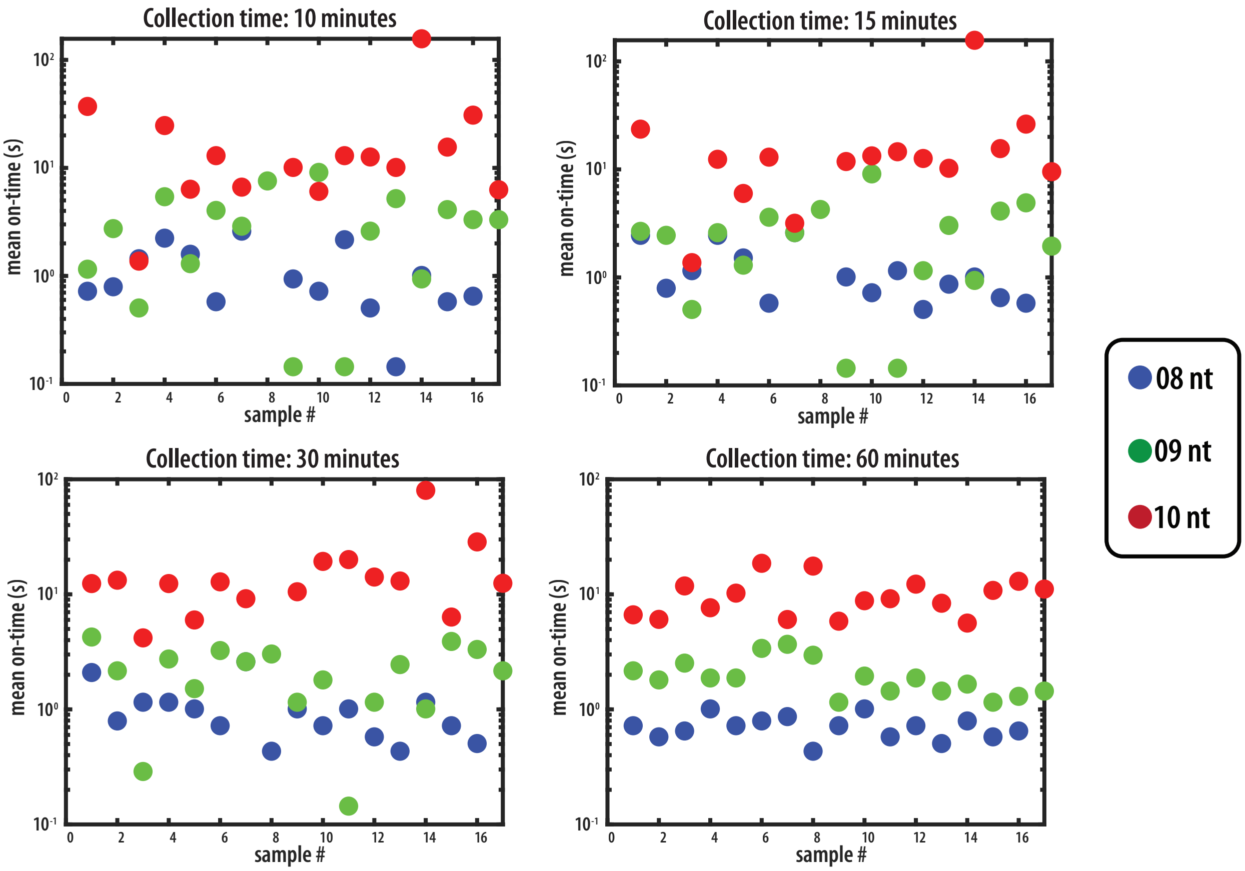

Figure S8: A control experiment which shows the decrease in the variance of the estimated mean on-time with increase in the total data collection time for 8, 9, and 10 nt devices. This is as expected from prior theoretical studies[12]

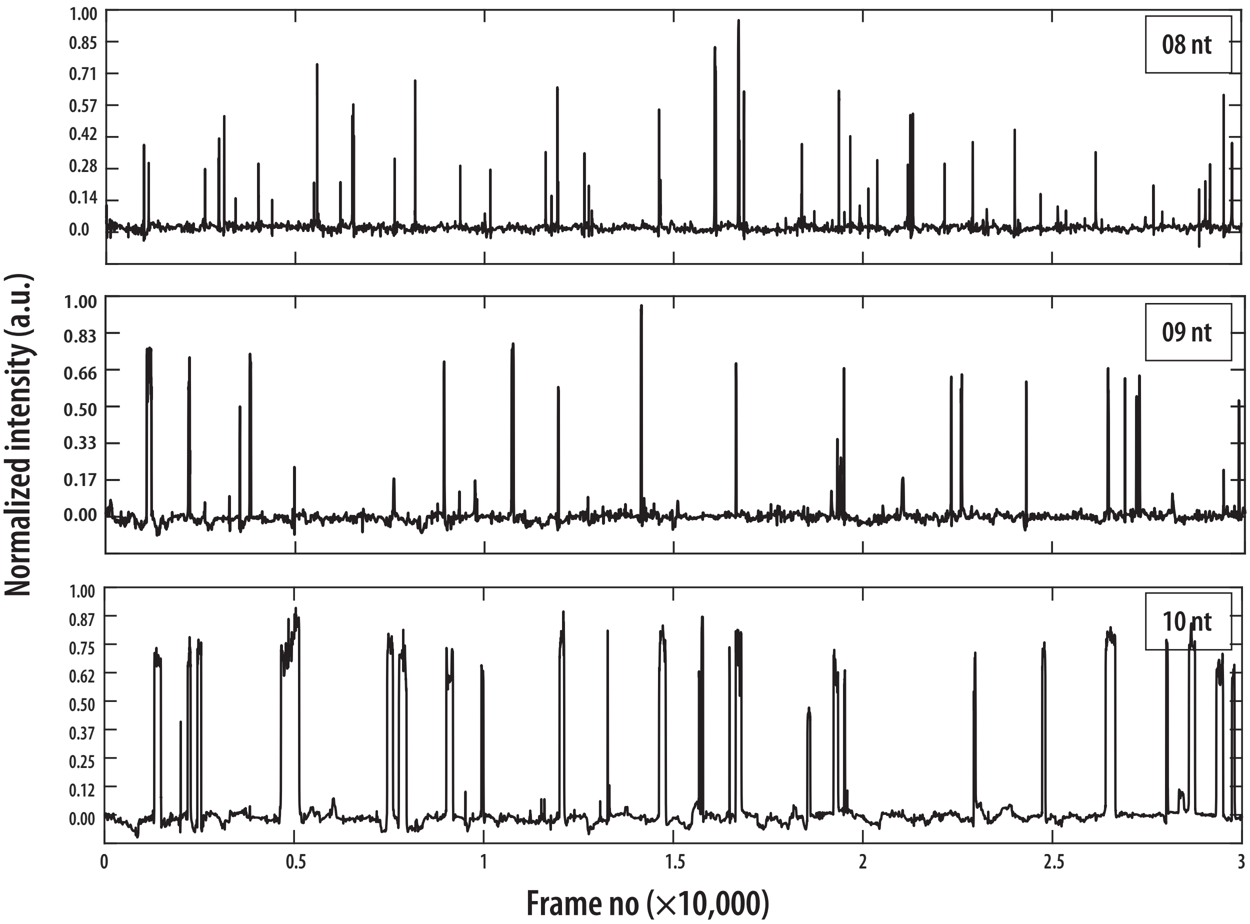

Figure S9: Exemplary time signals of the simple devices for visualization purposes. From top to bottom, each device has a single domain with lengths 08 nt, 09 nt and 10 nt respectively.

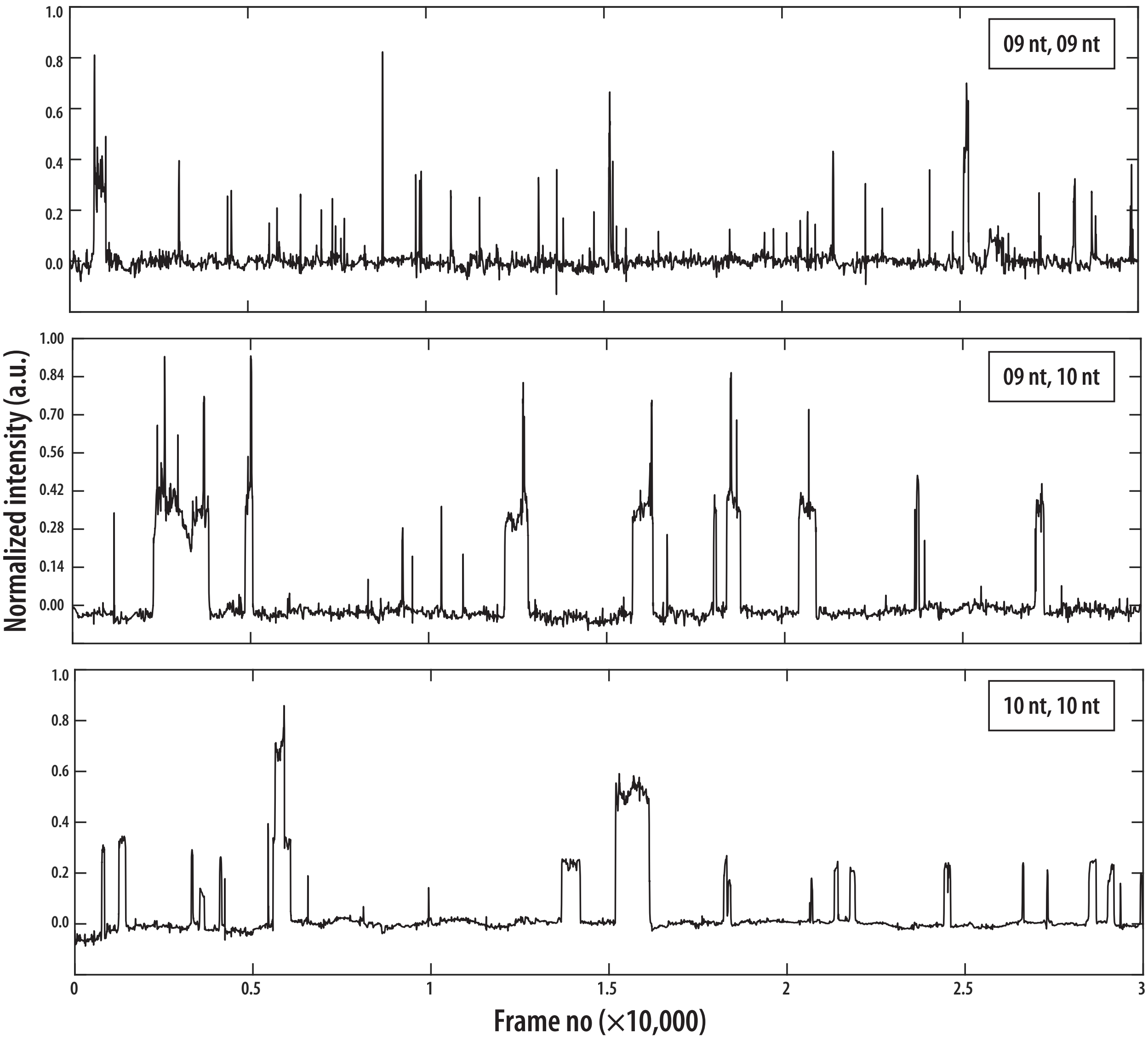

Figure S10: Exemplary time signals of the double-domain devices for visualization purposes. From top to bottom, each device has two domains with lengths 9-9 nt, 9-10 nt and 10-10 nt respectively.

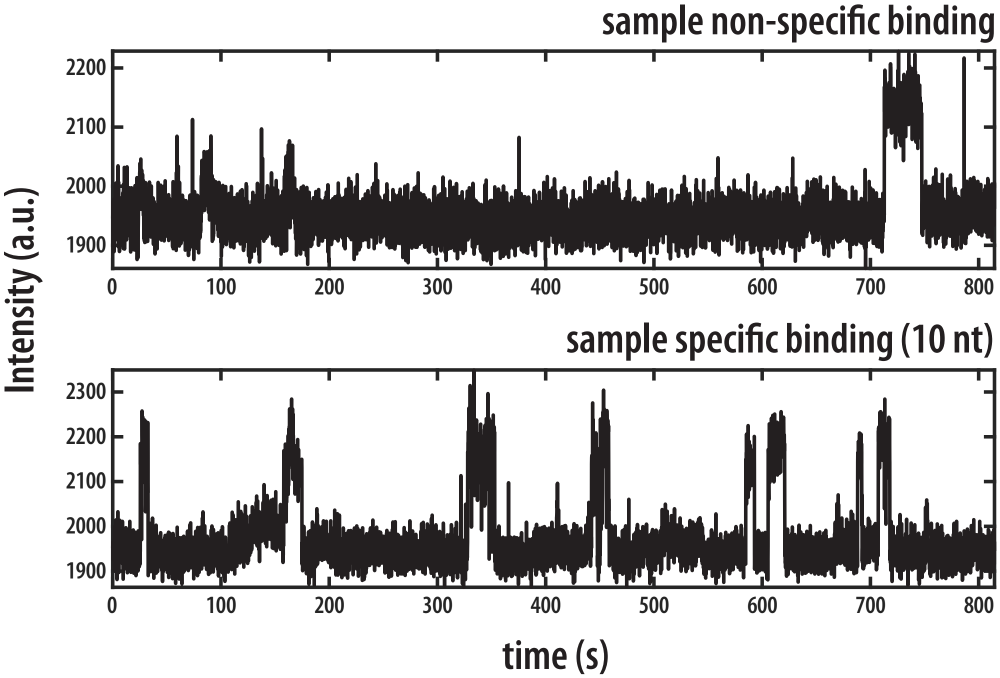

Figure S11: A typical temporal barcode of the nonspecific localization and a device localization. If data acquisition times are long, the difference in the number of on-peaks between signal and nonspecific localization is distinguishable. These nonspecific signals are removed using a software-based peak filter.

References

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

(2) Dai, M.; Jungmann, R.; Yin, P. *Nat. Nanotechnol.* **2016**, *11*, 798.

(3) Dai, M. *3D DNA Nanostructure*; Springer, 2017; pp 185–202.

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

(5) Wang, Y.; Schnitzbauer, J.; Hu, Z.; Li, X.; Cheng, Y.; Huang, Z.-L.; Huang, B. *Opt. Express* **2014**, *22*, 15982–15991.

(6) Ovesnỳ, M.; Křížek, P.; Borkovec, J.; Švindrych, Z.; Hagen, G. M. *Bioinformatics* **2014**, *30*, 2389–2390.

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

(8) Pirchi, M.; Tsukanov, R.; Khamis, R.; Tomov, T. E.; Berger, Y.; Khara, D. C.; Volkov, H.; Haran, G.; Nir, E. *J. Phys. Chem. B* **2016**, *120*, 13065–13075.

(9) Speicher, N. Oligo synthesis: Why IDT leads the oligo industry. `https://bit.ly/2CsXsET`, 2015; [Accessed: 2019-03-13].

(10) Lambert, T. tlambert03/FPbase. 2018; `https://github.com/tlambert03/FPbase`.

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

(12) 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.