Office of Naval Research (ONR)

Research Performance Final Report

ONR Project Title: DNA-based technologies for reading and writing large-scale molecular patterns with nanoscale-precision

ONR award number: N00014-18-1-2549

PI: Peng Yin

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

Distribution Statement

DISTRIBUTION A. Approved for public release: distribution unlimited.

Accomplishments

What were the major goals and objectives of the project?

Throughout the course of the grant period, we have developed a set of DNA-based technologies towards an over-arching goal of reading and writing large scale molecular patterns with molecular precision.

We proposed and implemented complementary strategies towards the technology development
(1) Direct nanoscale labeling and visualization of targets in user-prescribed positions
Direct manipulation and visualization of nanoscale targets is a fundamental challenge in nanotechnology. We have therefore developed a suite a both nanoscale labeling and visualization DNA-based labeling and imaging technologies that can be used both *in vitro* and *in situ*.

(2) Autonomous recording of nanoscale geometries using DNA-based proximity recorders.
Highly complex geometries at the nanoscale requires a fundamentally different approach to visualize. We have therefore developed technologies that can record nanoscale distances with autonomous DNA recorders.

We set out to accomplish these goals using a multi-pronged approach: Scaling up Action-PAINT labeling and developing combinatorial targeting. (Aim 1), Scaling up autonomous proximity recorders (Aim 2), Large scale patterning with nm-to-cm addressability (Aim 3).

What was accomplished towards achieving these goals?

Aim 1: Scaling up Action-PAINT labeling and developing combinatorial targeting.

Action-PAINT

A key development towards nanoscale labeling is our "Action-PAINT" technology, a novel super-resolution labeling technology platform which allows users to direct a DNA barcode to a region with 30 nm resolution, thereby surpassing a fundamental limit imposed by the diffraction limit of light (~200 nm). The work is now published in **Nature Chemistry** [1].

Software

Custom MATLAB laser control and real-time image analysis software were written and tested to work robustly in an experimental setting.

Nanoscale In-vitro labeling

DNA origami test structures were utilized as resolution benchmarks to assay the feasibility of Action-PAINT. Fast photocrosslinking DNA imaging oligos have been developed and a sequence library has been produced which can be used in combination with DNA-PAINT super-resolution imaging. Current testing has demonstrated that Action-PAINT can label single DNA strands with

30 nm spatial discrimination, demonstrating a first implementation of user-defined nanoscale labeling.

Nanoscale In-situ labeling
The features of Action-PAINT labeling, whereby a user can perform a super-resolved labeling workflow after imaging, would be of great value towards labeling specific protein targets in cellular or tissue samples, due to the lack of *a priori* knowledge of the macroscale structural organization of biomolecular complexes in cells. To this end we have demonstrated user-defined labeling of microtubule samples in fixed cellular samples.

Biocompatibility of assembled nanostructures
Unmodified DNA nanostructures would not be usable in an organic environment due to the presence of nucleases and other denaturants in cellular samples. Novel DNA synthesis and assembly techniques have been tested to assemble DNA nanostructures that have been found to be resistant to nuclease degradation and chemical denaturants and have also been found to be compatible with cell culture media. Current work is published in **Angewandte** [2].

Light-Seq
Based on our previously published Action-PAINT technology, we have scaled up labeling to target and barcode the whole transcriptome of biological targets *in situ* in an unbiased manner, parallelized labeling regions with a digital micromirror device (DMD), and applied the technology in the field of spatial transcriptomics to measure differences in gene expression across the spatial dimension in tissue samples. The technology is termed "Light-Seq" and is now published in **Nature Methods** [3].

Parallelized labeling of Action-PAINT with DMD arrays: DNA barcodes have been optimized and tested to label regions on a slide. An automated system of photomask creation for the DMD and illumination has been implemented that can direct a barcoded piece of DNA to any arbitrary position (Appendix 3, Figure 2). A manuscript is in preparation to disseminate the method (See publications under products, appendix 3).

Combinatorial Labeling strategy: We developed and are currently validating a combinatorial labeling strategy in order to scale the number of uniquely addressable features. The basic strategy involves the use of concatenating barcode strands to create a unique barcodes sequence. This strategy can scale the number of addressable features as M^N, where M is the size of the barcode library and N is the number of exchanges (Appendix 3, Figure 1). Currently we have a cyclical barcoding workflow that is integrated with a series of patterned illumination profiles on the DMD to independently and in parallel label >100 positions (Appendix 3, Figure 5)

Multi-species orthogonal labeling: We expanded our labeling strategy to include multiple orthogonal DNA species. Current system has demonstrated three cycle labeling of three orthogonal fluorescently labeled DNA barcode strands. Three photomasks where illuminated over three cycles to construct an image of a Penrose triangle (Appendix 3, Figure 4). This demonstration validates both the sequence and spatial specificity of the method.

Automation of labeling workflow: A high-throughput labeling workflow was developed to increase the throughput of nanoscale labeling. Addressing different regions with unique labeling barcodes

typically requires multiple rounds of buffer exchanges and labeling cycles, thus large portions of the workflow can be automated in order to increase throughput and reduce human error.

Software for feature detection: An automated feature detection and UV labeling set up was tested to create a hands-off system that can automate labeling of multiple features across multiple fields of view. Current system can achieve micron scale feature size labeling across millimeter scale fields of view (Appendix 3, Figure 3,5).

Automated fluidic exchange: Two different on-microscope fluid exchange systems were tested with our labeling workflow in conjunction with macro scripts that coordinate the communication between our microscope and the fluidic systems. By automating the previously manual fluid exchange steps (DNA hybridization and washing) and eliminating the need to remove the sample from the microscope, the throughput of the barcoding workflow was significantly increased. A full automated workflow of image, label, exchange was achieved on-scope for 20+ labeling cycles.

Aim 2: Scaling up autonomous proximity recorders

SABER: signal amplification by exchange reaction

Building on our previously developed Primer Exchange Reaction (PER), an autonomous signal amplification technology, we have expanded PER to be applied *in situ* to amplify imaging signals of both chromatin and protein targets.

In-situ amplifier of chromatin FISH experiments

PER amplification has been applied for multi-color chromatin imaging. By autonomously extending and copying a DNA docking site for a fluorescent DNA imager, the end effect is a signal amplification of the desired binding target. Current work has been published in **Nature Methods** [4]

In-situ amplifier of protein targets

A similar principle of PER amplification of DNA docking sites has been applied to protein antibodies labeled with DNA strands. Initial experiments have provided a list of validated antibodies and applied towards a 10-color imaging experiment of different protein targets in whole tissue samples. Current work has been published in **Nature Biotechnology** [5].

DNA Nanoscopy

Toward our goal of creating high throughput autonomous molecular proximity recorders, we further developed two methods (molecular rulers and molecular crawlers) based on our previously developed polymerase-driven Primer Exchange Reaction (PER) and Autocyclic Proximity Reaction (APR) synthesis method. Both of these technologies have now been shown to successfully measure and re-construct nanoscale architectures - see below and Appendices 1 and 2.

Autonomous molecular swarms for relative position information retrieval (molecular crawlers): A set of DNA based molecular agents has been developed that are designed for a molecular recording scheme for inspecting and reconstructing spatial arrangement of molecular landscapes. For the past year, the focus has been on realizing single-molecule level resolution based on unique barcoding and next-gen sequencing. Proof-of-concept experiments have been demonstrated on

pre-fabricated nanostructures with set spacing as well as in-situ tubulin targets (Appendix 1, figure 3). A manuscript is in preparation (See publications under products and Appendix 1).

Autonomous reconstruction of molecular landscapes (molecular rulers): A DNA nanodevice has been tested that can autonomously measure molecular scale pairwise distances and reconstruct spatial geometries using Next Generation Sequencing (NGS). Several complex geometries have been reconstructed using the molecular ruler system, including donut shapes and letters at the nanoscale. A manuscript has been posted on *bioRxiv* (See publications under products and Appendix 2)

Aim 3: Large scale patterning with nm-to-cm addressability

By interlacing a patterned illumination profile with moving the stage across multiple fields of view, we proposed to perform small feature sized labeling across large length scales. We have now successfully performed micron scale feature size labeling across mm length scales by focusing individual DMD mirrors into a dot array (Appendix 3, Figure 3). The current chip size of the DMD is focused onto a ~0.5 mm sized area to achieve its micron scale feature size and can be tiled across a 2x2 area to approach the mm length scale. Ultimately, we seek to combine this large scale patterning with our combinatorial nucleic acid assembly method (Aim 1 above) to have unprecedented control over nucleic acid labeling and positioning on arbitrary substrates.

What opportunities for training and professional development did the project provide?

Signal amplification via PER has already received great interest and application in multiplexed fluorescent tissue imaging. Several collaborations and have already been started with various medical research facilities at Harvard Medical School and Brigham and Women's hospital. An annual training session for PER is currently being developed to allow other labs to independently implement the PER imaging protocol. Signal amplification is now an integral part of the NIH Hubmap initiative and CZI Human Cell Atlas initiatives. It has been widely disseminated to all member groups of Hubmap.

Action-PAINT has been recently published and has been well-received. Several labs have reached out and expressed interest in applying the same crosslinking chemistry for research areas in DNA based data storage and RNA targeting.

Light-Seq has received wide interest upon publication. An online set of protocols as well as a series of Jupyter notebooks used in a Light-Seq "Zoom workshop" has already been presented to interested parties. The workshop will be restarted next year upon further interest. Two research technicians have been trained in the Light-Seq workflow. All online resources can be found at lightseq.io

How were the results disseminated to communities of interest?

Talks

Peng Yin. Clinical Pathology Conference, Brigham and Women's Hospital, Boston, MA, May 28th, 2019.

Peng Yin. Department of Pharmacology and Chemical Biology, Baylor College of Medicine, Houston, TX, April 23rd, 2019.

Peng Yin. Nanoscale Subgroup meeting, Biophysical Society Annual Meeting, Baltimore, MD, March 2nd, 2019

Peng Yin. Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, Feb. 26th, 2019.

Peng Yin. Department of Biomedical Engineering, Northwestern University, Evanston, IL, Jan. 24th, 2019.

Peng Yin. Department of Chemistry, Boston College, Chestnut Hill, MA, Nov. 28th, 2018

Peng Yin. Neurotechnology symposium at MIT, Cambridge, MA, Nov. 27th, 2018.

Peng Yin. Boston Biology and Biotechnology Association 25th Annual Symposium, Boston, MA, Oct. 6th, 2018.

Peng Yin. Wyss Institute International Symposium on Molecular Robotics, Boston, MA, Sep. 21st, 2018.

Peng Yin. Nanomedicines: From Fundamentals to Applications session, 256th American Chemical Society National Meeting, Boston, MA, Aug. 22nd, 2018.

Peng Yin. Nanoscience, Nanotechnology, and Beyond symposium, 256th American Chemical Society National Meeting, Boston, MA, Aug. 21st, 2018.

Peng Yin. Nucleic Acid-Based Sensors session, 256th American Chemical Society National Meeting, Boston, MA, Aug. 21st, 2018.

Sungwook Woo, Foundations of Nanoscience (FNANO). April 16th, 2019. Snowbird, UT.

Youngeun Kim, Transmutable Nanoparticles and Reconfigurable Nanoparticle Superlattices. 2019 93rd American Chemical Society Colloid & Surface Science Symposium, Atlanta GA *Keynote speaker*

Peng Yin, Department of Biomedical Engineering, University of Texas at Austin, May 13th, 2021 (Virtual)

Peng Yin, Workshop on Nucleic acids, synthetic biology and artificial life, Imperial College, London, Apr. 29th, 2021 (Virtual)

Peng Yin, European Molecular Biology Laboratory Biennial Conference, Heidelberg, Germany, November 15th-18th, 2020 (Virtual)

Peng Yin, Frontiers in Nanotechnology Virtual Mini-Conference, Northwestern University, Evanston, IL, July 22nd, 2020 (Virtual)

Nikhil Gopalkrishnan, Bio-compute club, Massachusetts Institute of Technology, Oct. 2020 (Virtual)

Nikhil Gopalkrishnan, ESCMID Conference on Coronavirus Disease (ECCVID), Sep. 2020 (Virtual)

Nikhil Gopalkrishnan, 26th International Conference on DNA Computing and Molecular Programming, Sep. 2020 (Virtual)

Nikhil Gopalkrishnan, SLAS Transformed, Jun. 2020 (Virtual)

Nikhil Gopalkrishnan, Chemical Biology, Brigham & Women's, Harvard Medical School and MIT, Jun 2020 (Virtual)

Nikhil Gopalkrishnan, Department of Systems Biology, Harvard Medical School, May 2020 (Virtual)

Nikhil Gopalkrishnan, Foundations of Nanoscience (18th annual conference), Apr. 2020 (Virtual)

Ninning Liu, Department of Systems Biology, Harvard Medical School, September 2022.

Ninning Liu, Wyss Institute Symposium, Harvard Medical School, November 2022.

Posters

Ninning Liu. Wyss Institute Annual Symposium. Nov 15, 2018. Boston, MA.

Ninning Liu. Department of Systems Biology at Harvard Medical School Retreat. June 5, 2019. Sebasco, ME.

Ninning Liu, Wyss Institute Annual Symposium. November 2022. Boston, MA. *Poster*

Youngeun Kim. Foundations of Nanoscience (FNANO). April 16th, 2019. Snowbird, UT.

Sungwook Woo, Wyss Institute Annual Symposium. Nov 15, 2018. Boston, MA.

Sungwook Woo, Poster, Wyss retreat, "Molecular Crawlers for Inspection and Reconstruction of Molecular Landscapes", 11/22/2019, Boston, MA

Youngeun Kim. "DNA Dendrimer Coated DNA Nanostructures" 2019 16th Annual Conference on Foundations of Nanoscience, Snowbird UT *Poster presenter*

Online Resources:
SABER-FISH technology: http://saber.fish/
Immuno-SABER Technology: http://immuno-saber.net/
Light-Seq: https://www.lightseq.io/

Online repositories:
SABER and Immuno-SABER software:
https://yin.hms.harvard.edu/SABER/SABER-FISH.html#resources
Light-Seq sequencing data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE208650
Light-Seq, one image set: https://ninning.github.io/wyss-retreat/
Light-Seq software: https://github.com/Harvard-MolSys-Lab/Light-Seq-Nature-Methods-2022

Technology Transfer

Work on this grant has led to the filing of 6 separate patent applications (see separate Patent Report).

Award Participants

- Peng Yin (PI)
- Youngeun Kim (Postdoc)
- Sungwook Woo (Postdoc)
- Ninning Liu (Postdoc)
- Fan Hong (Postdoc)
- Kuanwei Sheng (Postdoc)
- Nikhil Gopalkrishnan (Postdoc)
- Yan Yan (Postdoc)
- Swarup Day (Postdoc)
- Adam Yaseen (Technician)
- Jonathan Jordanides (Technician)
- Weidong Xu (Graduate Student)
- Gokul Gowri (Graduate Student)

Products

Published Articles

[1] N. Liu, M. Dai, S. K. Saka, and P. Yin, "Super-resolution labelling with Action-PAINT," *Nat. Chem.*, vol. 11, no. 11, pp. 1001–1008, Nov. 2019.
[2] Y. Kim and P. Yin, "Enhancing Biocompatible Stability of DNA Nanostructures Using Dendritic Oligonucleotides and Brick Motifs," *Angew. Chem. Int. Ed Engl.*, vol. 59, no. 2, pp. 700–703, Jan. 2020.
[3] J. Y. Kishi *et al.*, "Light-Seq: light-directed in situ barcoding of biomolecules in fixed cells and tissues for spatially indexed sequencing," *Nat. Methods*, vol. 19, no. 11, pp. 1393–1402, Nov. 2022.

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

[4] J. Y. Kishi *et al.*, "SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues," *Nat. Methods*, vol. 16, no. 6, pp. 533–544, Jun. 2019.
[5] S. K. Saka *et al.*, "Immuno-SABER enables highly multiplexed and amplified protein imaging in tissues," *Nat. Biotechnol.*, vol. 37, no. 9, pp. 1080–1090, Sep. 2019.

Preprint articles (see Appendix for drafts)

S. Woo, P. Yin. **Molecular swarm agents that directly interrogate molecular landscapes for information retrieval.** *In preparation.*

Key Words: DNA nanotechnology, molecular recorders
Abstract:
We report molecular swarm agents that directly interrogate molecular landscapes for information retrieval and quantitative analyses. Our method does not require purification of target species nor require samples to be immobilized on a substrate, while allowing a whole-sample analysis *in situ*. We describe our system as a group of 'agents', as our molecular species record information from the target molecules and report it back to us, without us having to directly examine the targets by either taking them out of the system or visualizing them. The use of agents allows the target molecules to be kept intact and the information to be sampled repeatedly. We depict the agents as a 'swarm', as they work independently of each other and operate on multiple targets at the same time, reporting back the collection of information.

N. Gopalkrishnan, S. Punthambaker, T. Schaus, G. Church, P. Yin. **A DNA nanoscope that identifies and precisely localizes over a hundred unique molecular features with nanometer accuracy.** *bioRxiv*. Posted 28 August 2020.

Key words: DNA nanotechnology, self-assembly, molecular recorders
Abstract:
Techniques that can both spatially map out molecular features and discriminate many targets would be highly valued for their utility in studying fundamental nanoscale processes. In spite of decades of development, no current technique can achieve both nanoscale resolution and discriminate hundreds of targets. Here, we report the development of a novel bottom-up technology that: (a) labels a sample with DNA barcodes, (b) measures pairwise-distances between labeled sites and writes them into DNA molecules, (c) reads the pairwise-distances by sequencing and (d) robustly integrates this noisy information to reveal the geometry of the underlying sample. We demonstrate our technology on DNA origami. We both spatially localized and uniquely identified over a hundred densely packed unique elements, some spaced just 6 nm apart. The bottom-up, sequencing-enabled mechanism of the DNA nanoscope is fundamentally different from top down imaging, and hence offers unique advantages.

Appendix 1

Molecular swarm agents that directly interrogate molecular landscapes for information retrieval

Introduction

Human curiosity has driven exploration of unknown realms across scales, and it has led to the development of instruments from microscopes to telescopes and from tweezers to spaceships. For territories that are hard for humans to directly reach, 'agents' were often sent. For example, Mars rovers were dispatched to the distant planet to collect various data and send them back to Earth. Rescue robots aid in searching survivors in disaster conditions where the task may be too dangerous for humans. In the Internet space, we have web crawlers to visit and index web pages that are overwhelmingly too many. Retrieving information from our agents is a crucial step that gives us new knowledge, important clues, or useful insights about the hard-to-reach target.

For exploring the molecular world, particularly for studies of biomolecular interactions, scientists have taken approaches of either taking the molecules out of their natural environment or just viewing the molecules from a 'distance' using microscopes. One representative class of methods of the first approach is the co-purification methods for proteins, where the interacting molecules are isolated together and subjected to analysis. Examples include yeast two-hybrid screening[] and co-immunoprecipitation[], followed by identification processes such as western blotting or mass spectrometry. In addition to these methods typically requiring high purity and high abundance of the target molecules, since the molecules are analyzed outside of the cellular environment, the physiological context is often missed, and the native interactions may not be fully preserved along the usually harsh purification processes. The second approach of viewing the molecules denotes microscopy, where molecules of interest are tagged and directly visualized for detection of colocalization of interacting molecules. In particular, recent developments of various super-resolution techniques[] allowed visualization of objects below the diffraction limit, even down to the levels of individual molecules in a densely packed environment[]. However, these methods require samples to be immobilized on a substrate, and typically suffer limited throughput as only a limited population within a field of view is examined at a time and post-processes are usually required.

Here we report molecular swarm agents that directly interrogate molecular landscapes for information retrieval and quantitative analyses. Our method does not require purification of target species nor require samples to be immobilized on a substrate, while allowing a whole-sample analysis *in situ*. We describe our system as a group of 'agents', as our molecular species record information from the target molecules and report it back to us, without us having to directly examine the targets by either

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

taking them out of the system or visualizing them. The use of agents allows the target molecules to be kept intact and the information to be sampled repeatedly. We depict the agents as a 'swarm', as they work independently of each other and operate on multiple targets at the same time, reporting back the collection of information.

For the agents, we use DNA along with polymerases. As a natural information-encoding molecule, DNA has proven to be an excellent resource for molecular engineering for nanoscale assemblies[] and dynamic functions, such as circuits[], walkers[] and machines[1–7]. With polymerases and other enzymes added in the toolset, DNA-based systems have been engineered to program walking mechanisms[8–12], reaction networks[13,14] and rotational motors[15], and to synthesize arbitrary DNA[16] for multiple functions. DNA-bound microparticles along with enzymes were used to create microscopic agents[17], but the system did not support molecular resolution nor information transfer to a researcher other than through fluorescence signals. Additionally, DNA sequencing-based spatial reconstruction methods based on amplicon diffusion[DNA microscopy] or amplicon colonies[Hogberg PNAS work] have been demonstrated or proposed, and these methods similarly exhibit micron-scale resolution.

DNA has also been used to study biomolecular interactions in proximity assays: typically, DNA-based probes are used to tag target molecules and the colocalization of the molecules allows generation of a DNA signature, which then gets amplified and analyzed. Examples of proximity assays include proximity ligation[] and extension[] assays (PLA, PEA, respectively), and more recent systems that use more sophisticated probe designs and readouts such as proximity PAINT imaging methods[] and proximity-dependent hybridization chain reactions (ProxHCR)[]. Most of these methods are 'destructive', however, in the sense that each probe can only be used once, often leaving dead spots and resulting in incomplete analyses (Supplementary Figure 1 (originally Figure 1)). We previously developed the auto-cycling proximity recording (APR) reaction[] to overcome this limitation, where the probes were 'non-destructive', allowing repeated recording of proximity relationships between different partners, hence enabling complete spatial organizations to be revealed after subsequent reconstruction processes. Additionally, a similar non-destructive approach based on renewable DNA probes through manual cycling, termed iterative proximity ligation, has been proposed[Ellington/Marcotte]. However, the measurements in these proximity assays are typically limited to only *pairwise* interactions, making direct detection and analysis of interactions involving multiple biomolecules challenging.

Our swarm agents use a unique molecular mechanism that allows direct examination of DNA-labeled molecular targets *in situ* across *multiple* targets in proximity, in an autonomous and catalytic manner. Our scheme generates a molecular entity that dynamically trails and grows along DNA probes, which we call a 'molecular crawler'. When finished crawling, the molecule serves as a 'record' that reflects the trajectory and contains information copied from the target-bound probes. We demonstrate two key capabilities for quantitative analysis with crawlers: counting the number of subunits in molecular complexes and detecting multivalent proximity interactions.

Quantitative understanding has been crucial in biological studies. For example, the three-state model[] for the operation of ATP synthase, which contributed significantly to our understanding of the machinery, had been proposed solely based on quantitative understanding of reaction products(check),

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

before the crystal structure was discovered[]; membrane proteins typically form clusters of certain sizes when initiating signal cascades, thus it is often important to find the critical size of clusters to understand signal pathways and mechanisms[]; some enzyme complexes called morpheeins exhibit different activities depending on the number of subunits[]. Therefore, the counting capability is expected to have a wide range of uses. In our scheme, the length of each DNA record produced represents the number of subunits visited by the crawler, hence the maximum length revealed through repeated catalytic recording rounds gives us the size of the complex. We demonstrate the counting capability for two model systems, one with streptavidin and the other with artificial complexes of programmed sizes implemented on DNA origami[18].

Our multivalent colocalization detection scheme utilizes the feature of our design that, since it generates concatenated molecular records after visiting the component molecules, detecting the presence of the records *in situ* directly reveals multivalent interactions. The records can be collected and further subjected to diverse post-processes such as gel analysis or sequencing. Previously, a PLA-based microscopy study showed the detection of colocalization of three protein species[19]; however, in addition to being a destructive and low-throughput method as described above, the PLA mechanism required multiple probe species to bind together in a stoichiometry-sensitive manner, limiting the scalability. Our DNA probes need not bind together at the same time to detect multiple interactions, as each step only requires two proximal partners to interact at a time. As each step is local and independent of previous steps regardless of the total number of components, our mechanism is highly scalable and in principle boundless in terms of the number of detectable interactions. We show multi-step records of size up to 11, limited only by the size of the track. In addition, since the recording mechanism is non-destructive and catalytic, multiple rounds of recording can occur on the same target, hence amplifying signals that might otherwise remain negligible. We apply the mechanism to detecting colocalization of three protein species at the microtubule growing end *in situ*, preserving the physiological context. We distinguish different colocalization patterns depending on three different drug treatment states. We present gel results of collected records, as well as confirm the generated records *in situ* in the physiological condition by fluorescence microscopy.

Results

Design of the crawler system. Each of the probes, or the track molecules along which a crawler follows and grows, is composed of two domains: a primer-binding domain and a copy-and-release (CR) domain. The primer-binding domain takes in a primer which then gets extended by a polymerase into the CR domain. The CR domain is a double-stranded motif that allows first copying of an arbitrary sequence sequestered in the domain onto a primer and then spontaneous release of the copied segment under an isothermal condition [APR, PER]. The bottom of the CR domain is met with a 'stopper' point beyond which the polymerase cannot proceed. The stopper is encoded by a noncanonical base pair iso-C/G, or a carbon chain spacer. The CR domain in the crawler system is further divided into two subdomains: a barcode domain and a primer-encoding domain. The barcode domain contains a sequence that uniquely identifies the specific probe. The domain can also be used as a spacer domain, depending on the purpose, e.g., in

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

the counting scheme described later. The primer-encoding domain holds a sequence that encodes a primer for a subsequent step. The structure of a probe and basic operation mechanism are described in Figure 1a. The top row depicts the unit operation on a single probe. The reaction is initiated by the binding of a primer (strand 'a') from the solution onto the primer-binding domain (a*) of the probe. The next step is the elongation of the primer by a strand-displacing polymerase (typically *Bst*, New England Biolabs) along the template. Since the newly synthesized part shares the same sequence as the dangling side of the CR domain (domains '1-b'), the new part and the dangling segment can undergo a random walk branch migration process[]. If the original segment in the CR domain displaces the newly synthesized part, a new primer for the next reaction is exposed (domain 'b').

The new primer now initiates the next reaction (middle row of Figure 1a). Since the growing crawler is still anchored on the first probe, the new primer only acts locally on probes in proximity. When the new primer binds the primer-binding domain of the next probe through complementarity, the same operation is repeated. The outcome is a crawler with its body spanning across two probes. The primer-binding domains of the second probe or thereafter (e.g., domain 'b*' of the second probe) are protected against elongation by a polymerase, to prevent its own extension which would result in permanent modification of the probe; such protection is achieved by incorporating inverted-dT, a non-extendable base, at the 3'-end of the strand. When three probes with primer sequences designed to match in series are in proximity (as in the bottom row of Figure 1a), the reaction yields an extended crawler spanning across three probes. A variation of this mechanism, where the numbered domains ('1', '2', and '3') were designed as single-stranded DNA, was also used in later parts of this study (see Supplementary Figure#).

One of the key properties of the crawler system is that the recording process is non-destructive to the probes, leaving the state of each probe unchanged after each round, hence allowing catalytic, repeated recording. One central mechanism that allows this property is the autonomous release of a record, and it is enabled by another DNA species in solution, which we call a "release primer" (strand 'd*' in Figure 1a). When a crawler reaches a certain probe that encodes the complement of the release primer (e.g., domain 'd' in probe numbered 3 in Figure 1a), and the complement in the newly synthesized domain becomes exposed, the release primer binds it and a polymerase can extend the release primer all the way to the end (a*), spontaneously releasing the double-stranded record into the solution and returning the probes to the original state. This process also naturally allows retrieval of the records through collection of the supernatant solution.

Proof-of-concept. To demonstrate the basic mechanism of the molecular crawlers, we built a three-point track with a triangular alignment on a DNA origami platform. Figure 1b depicts a schematic of the design (top) and the strand diagram of a crawler after crawling over the three probes (bottom). DNA origami were deposited on a mica substrate to prevent damage by the polymerase [APR] and to allow atomic force microscope (AFM) characterization. Figure 1c shows models and averaged AFM images of an origami rectangle before (i, ii) and after (iii, iv) a crawler has trailed along the three probes. Before recording, the probes appear as dots because the sweeping action of an AFM tip pushes around the tethered probes and can only capture faint images of the anchor points. After the recording reaction of about 1 hr, but without

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

the release primer added, the crawler now connects and holds the three probes together as shown in Figure 1b bottom, and thus appear accordingly in the AFM image (Figure 1c (iv)). The fully grown crawler record reaches a length of 100 nucleotides (nt). When the records were retrieved by including the release primer during the reaction, amplified by PCR and run on a denaturing gel, the final records appear at the expected length range (Figure 1d). The sequence was also confirmed using Sanger sequencing (Figure 1e).

We further demonstrated the scalability of the crawling action on a prescribed track, where a crawler was allowed to move nine steps in total and was analyzed by AFM imaging. For multi-step tracks, the scalability of a crawling reaction would primarily be determined by the number of orthogonal primers available, if each step were to be uniquely prescribed. However, since the track sites are spatially restricted, not all of the primers have to be fully orthogonal to one another; i.e., for sites that are far enough in space to interfere, the primers in those sites do not need to be strictly orthogonal. Hence, repeated alignments of a set of distinct probes along a desired path can allow multiple steps in a prescribed manner. Figure 1f shows such an example where a set of three kinds of probes (each marked 'b-c', 'c-d', and 'd-b') are placed along the boundary of a shape in order and in a repeated fashion. After the initiating primer ('a') lands on the start site ('a-b') and crawls along the path until it hits a 'd-b' probe, it then connects to a nearby 'b-c' probe again and continues crawling until it finishes the path. Figures 1f (i) and (ii) show representative images of the rectangle origami before and after the reaction, respectively, and Figure 1f (iii) is an average of the images taken after the reaction.

Random crawling and counting. For the crawler system to allow wide applications, it is desired to have the ability to randomly move around and examine unknown targets, rather than in a fully prescribed fashion as shown in the previous section. We achieved a random crawling behavior by introducing a single type of 'universal' probe, which allows initiation, crawling and release of a crawler at any site. Such a universal probe can be designed by incorporating a 'tandem' primer-binding domain and 'redundant' primers. A tandem primer-binding domain has two primer-binding domains, e.g., 'a*' and 'b*', concatenated together (Figure 2a), such that it can bind either primer 'a' or 'b'. Redundant primers denote the arrangement of primers such that the incoming primer and the outgoing primer on a single probe are the same (the domains marked 'b' in the example shown in Figure 2a). When primer 'a' binds the 'a*' section of the tandem primer-binding domain and gets extended by a polymerase, a new domain 'b' is generated at the bottom of the probe. The newly created domain 'b' can then act as a primer for a crawling reaction to a nearby probe by binding the 'b*' segment of the tandem primer-binding domain of the next probe, thereby allowing continuation of the crawling reaction. Since each universal probe has the 'a*' domain, the crawling reaction can initiate at *any* site. Since 'b' is the connecting primer that allows crawling between probes, the crawler can proceed to *any* site in proximity. Using strand 'b*' as a release primer, the crawler can be released at *any* site as well. The combination of these rules allows a truly random crawling mechanism. The molecular detail is shown in Figure 2b.

This random crawling principle naturally allows a mechanism for 'counting' – counting the number of subunits within a molecular complex, because the longest possible record generated from a molecular complex reflects the maximum number of steps that can be taken within the complex and in turn the

number of subunits. We first demonstrate this capability by using a model system based on streptavidin. Streptavidin is a well-known tetrameric protein, where each monomer contains one binding site for biotin. We incorporate a universal probe that contains a tandem primer binding domain (b*-a*) and encodes 'b' as a new primer (collectively, denoted 'a-b-b') at each biotin pocket through a biotinylated DNA strand as shown in Figure 2a. An initiating primer 'a' can start a reaction at any of the four sites, then crawl to any of the nearby sites if available or be released by a release primer ('b*') at any site. This allows four kinds of records to be generated, each with a distinct length. The shortest one is a 'half record' from only one of the probes, released by a release primer before the crawler was able to crawl to a nearby probe. The longest one is a 'full record' generated by a crawler that visited all four probes on a single tetramer. There are two intermediate-length records that are created when a crawler visited two or three subunits on a tetramer before getting released. The four kinds of records with distinct lengths show up as distinct bands on a gel as shown in Figure 2c (left lane, marked '+SA'). In the absence of streptavidin, the crawler can only count up to '1', as the probes are freely floating in solution without forming a complex, as shown in the right lane (marked '-SA'). The shortest record that corresponds to count one is made from a single probe whether or not it is bound to a complex, as expected.

We further demonstrate the counting capability in a more programmable fashion, by using DNA origami to create artificial molecular complexes with tunable size. Since recording on surface-bound DNA origami requires PCR amplification due to the low concentration, we assigned special probes on the start site ('a-b') and the finish site ('b-d') such that we can amplify the final products with primers 'a' and 'd*'. Between the start and finish sites, we set four variable positions. For these positions, we used probes with redundant primers ('b-b') and we vary inclusion of these probes (e.g., from none to all four) to change the size of the complex from two to six in total, including the start and finish sites (Figure 2d). Figure 2e describes the reaction mechanism and Figure 2f shows the gel data showing distinct bands for the counts for each complex. Since we used 'd*' as the release primer and the records were amplified, a 'half record' from a single probe does not appear in the data unlike in the streptavidin case. The first, lowest molecular weight band comes from a crawler connecting the start site and the finish site, and, as one variable subunit gets added, one more band appears above, thus the number of bands plus one indicates the size of the complex. This example also highlights the power of crawlers as molecular 'agents', as the crawlers take the measurements from a low concentration sample then amplify the results for easy readout.

Multivalent interaction detection. The unique property of molecular crawlers that they create concatenated records from proximal interactions regardless of the number of constituent components makes them a powerful tool for detection of multivalent proximal interactions. Multivalent interactions can be detected directly *in situ* by observing the generation of the corresponding records. The records can also be collected and further subjected to post-processes such as PCR amplification, gel analysis or sequencing.

We demonstrate this capability inside the cell, for a trivalent colocalization interaction of alpha tubulin, beta tubulin and end-binding protein (EB1) at microtubule growing ends. While alpha and beta tubulins form heterodimers that construct microtubules and thus are found ubiquitously along

microtubules and inside the cell, EB1 only interacts with the growing end of a microtubule, involving in stabilization of the growing end and regulation of other proteins[]. If a microtubule experiences catastrophic disassembly or otherwise goes out of the growing phase, EB1 dissociates from the microtubule. We set out to detect the colocalization of these three proteins at the growing end of microtubules using the crawler system. We target each protein using a respectively specific antibody labeled with a DNA anchor, to which a crawler probe specific to the protein was designed to bind (details in Supplementary Information). The probes are designed such that recording starts when the initiating primer 'a' binds the 'a*' domain of the 'a-b' probe on alpha tubulin, then proceeds to the 'b-c' probe on beta tubulin, and then on to the 'c-d' probe on EB1 (Figure 3a). In the presence of a release primer 'd*', the trivalent records—and only they—can be collected and analyzed. In separate recording rounds, if different primer sets are used, the presence of each protein monomer can be independently detected: e.g., 'a' and 'b*' can generate a short monovalent record from the probe bound to alpha tubulin, 'b' and 'c*' from beta tubulin, and 'c' and 'd*' from EB1, respectively. Depending on the growth state of the cell, the relative levels of the trivalent and monovalent records would show up differently; e.g., in an actively growing state, the level of trivalent records is expected to increase.

To illustrate the ability of the crawler system to distinguish different growth states of cells, we designed a set of experiments, where we divide cells into three population groups and treat them in parallel. The first group, (1), was treated with growth medium, so that a majority of the cells are in an actively growing state. We expect to see both trivalent and monovalent records to be significant in this state. The second group, (2), was treated the same way as group (1), then subsequently was treated with nocodazole, a drug that promotes microtubule disassembly[], for 1 hour at 37°C. For this group, due to the effect of the drug, the level of trivalent records is expected to drop; however, since the monomers, both the tubulins and EB1, are still present within the cells, the level of monovalent records is expected to remain unchanged. The third group, (3), was treated the same way as group (2), then additionally was treated again with growth medium for 1 hour such that the cells recover from the drug and microtubules can form again. In this state, the level of trivalent records is expected to return to a significant level, while the monovalent records would still be comparable. After recording for trivalent and monovalent records for each group, and PCR amplification, the results shown in gel are highly consistent with these expectations (Figure 3b). While the level of monovalent records for alpha tubulin and EB1 remained quite the same throughout the steps, the level of trivalent records dropped to ~28.4% for group (2) and recovered to ~93.9% after the regrowth treatment for group (3), relative to the gel band intensity of the trivalent records of group (1).

While the gel results successfully proved the ability of the crawler system to detect multivalent protein interactions and to distinguish different growth states of cells, we further confirmed the distinction by directly observing the crawler-generated records *in situ*. As the trivalent records—and only they—will expose the final primer 'd', if we add, instead of the regular release primer 'd*', a fluorescently labeled strand that will bind 'd' but will not be extended by a polymerase by incorporating a non-extendible base, then we will be able to spatially localize the trivalent records that were just created *in situ*. Note that we used a fluorescent strand 'd*-3*' for increased stability (Figure 3c (1) inset). The fluorescence microscopy results are as expected (Figure 3c). For group (1) cells, which are in an actively growing state, we observed fluorescence signals strongly appearing near and along the periphery of cells, whereas for group (2) cells,

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

where the drug disassembled microtubules, hence there is no growing microtubule ends, fluorescence signals were barely detectable or remained at the background level. After the recovery treatment, group (3) cells restored the growth state and exhibited active growing ends again near and along the periphery of cells. We took intensity profiles of the fluorescence signals along the axis from the center of the cell nucleus to the boundary (an example trace shown as a yellow line in Figure 3c (1)). We then normalized those profiles with the distance to the edge and superimposed them. Those plots are shown in Figure 3d for each group, along with an average profile (orange line). While, in group (1), a big bump near the periphery of cells is observed, in group (2), the signal is flat while also being very low (~OO% of the basal line of group (1)). In group (3), we clearly see the bump being restored near the periphery.

Discussion

... (in progress)

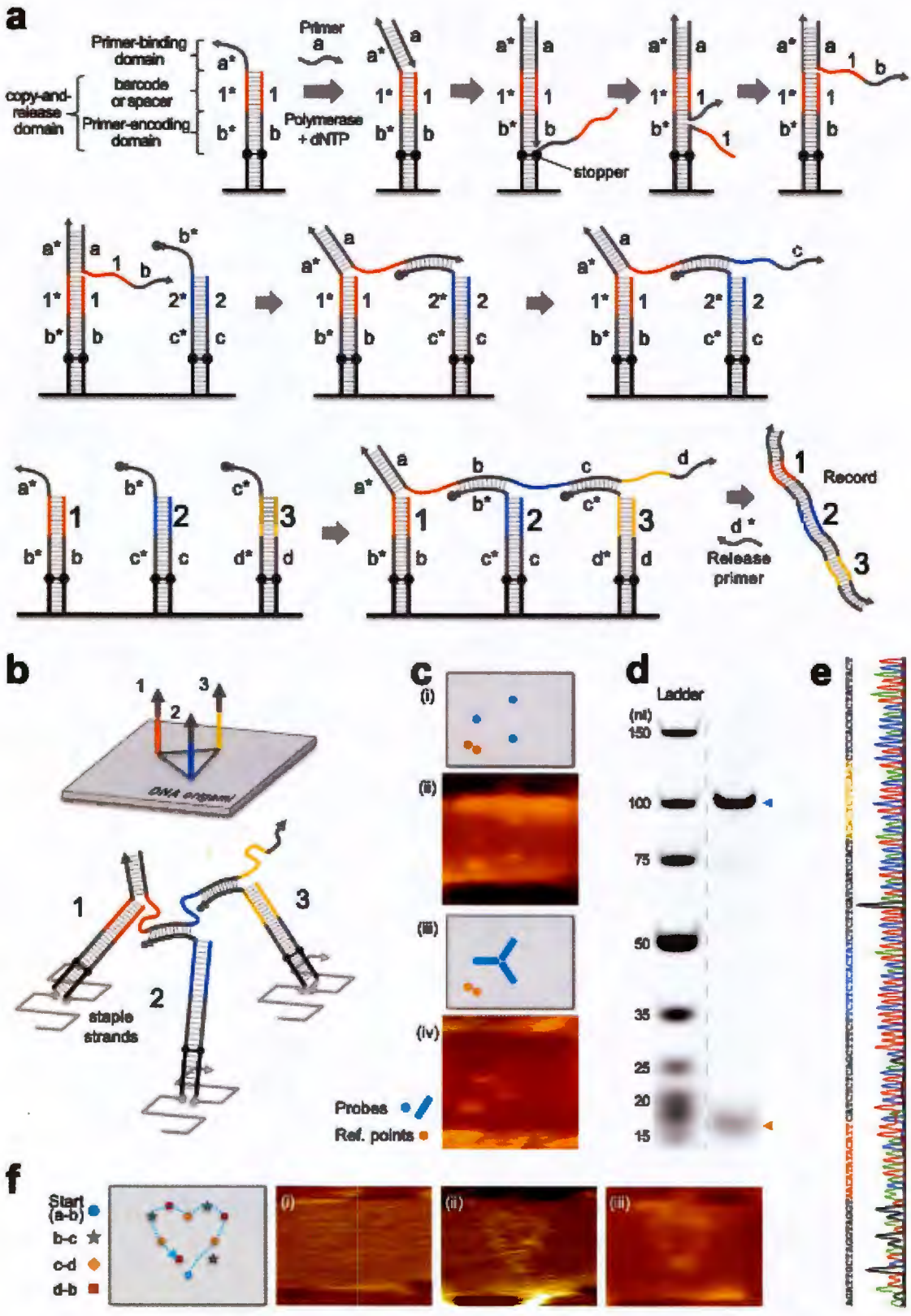
References

1. Yurke, B., Turberfield, A. J., Mills, A. P., Jr., Simmel, F. C. & Neumann, J. L. A DNA-fuelled molecular machine made of DNA. *Nature* **406**, 605–608 (2000).
2. Yan, H., Zhang, X. P., Shen, Z. Y. & Seeman, N. C. A robust DNA mechanical device controlled by hybridization topology. *Nature* **415**, 62–65 (2002).
3. Ding, B. & Seeman, N. C. Operation of a DNA Robot Arm Inserted into a 2D DNA Crystalline Substrate. *Science* **314**, 1583–1585 (2006).
4. Chakraborty, B., Sha, R. & Seeman, N. C. A DNA-based nanomechanical device with three robust states. *Proceedings of the National Academy of Sciences* **105**, 17245–17249 (2008).
5. Gu, H., Chao, J., Xiao, S.-J. & Seeman, N. C. A proximity-based programmable DNA nanoscale assembly line. *Nature* **465**, 202–205 (2010).
6. Douglas, S. M., Bachelet, I. & Church, G. M. A Logic-Gated Nanorobot for Targeted Transport of Molecular Payloads. *Science* **335**, 831–834 (2012).
7. Thubagere, A. J. *et al.* A cargo-sorting DNA robot. *Science* **357**, eaan6558 (2017).
8. Yin, P., Yan, H., Daniell, X. G., Turberfield, A. J. & Reif, J. H. A Unidirectional DNA Walker That Moves Autonomously along a Track. *Angew. Chem. Int. Ed.* **43**, 4906–4911 (2004).
9. Tian, Y., He, Y., Chen, Y., Yin, P. & Mao, C. A DNAzyme That Walks Processively and Autonomously along a One-Dimensional Track. *Angewandte Chemie International Edition* **44**, 4355–4358 (2005).
10. Bath, J., Green, S. J. & Turberfield, A. J. A Free-Running DNA Motor Powered by a Nicking Enzyme. *Angew. Chem. Int. Ed.* **44**, 4358–4361 (2005).
11. Lund, K. *et al.* Molecular robots guided by prescriptive landscapes. *Nature* **465**, 206–210 (2010).
12. Qu, X. *et al.* An Exonuclease III-Powered, On-Particle Stochastic DNA Walker. *Angew. Chem. Int. Ed.* 4 (2017).

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

13. Montagne, K., Plasson, R., Sakai, Y., Fujii, T. & Rondelez, Y. Programming an *in vitro* DNA oscillator using a molecular networking strategy. *Mol Syst Biol* **7**, 466 (2011).
14. Baccouche, A., Montagne, K., Padirac, A., Fujii, T. & Rondelez, Y. Dynamic DNA-toolbox reaction circuits: A walkthrough. *Methods* **67**, 234–249 (2014).
15. Valero, J., Pal, N., Dhakal, S., Walter, N. G. & Famulok, M. A bio-hybrid DNA rotor–stator nanoengine that moves along predefined tracks. *Nature Nanotech* **13**, 496–503 (2018).
16. Kishi, J. Y., Schaus, T. E., Gopalkrishnan, N., Xuan, F. & Yin, P. Programmable autonomous synthesis of single-stranded DNA. *Nature Chemistry* **10**, 155–164 (2017).
17. Gines, G. *et al.* Microscopic agents programmed by DNA circuits. *Nature Nanotech* **12**, 351–359 (2017).
18. Rothemund, P. W. K. Folding DNA to create nanoscale shapes and patterns. *Nature* **440**, 297–302 (2006).
19. Söderberg, O. *et al.* Direct observation of individual endogenous protein complexes in situ by proximity ligation. *Nat Methods* **3**, 995–1000 (2006).

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025



OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

Figure 1:

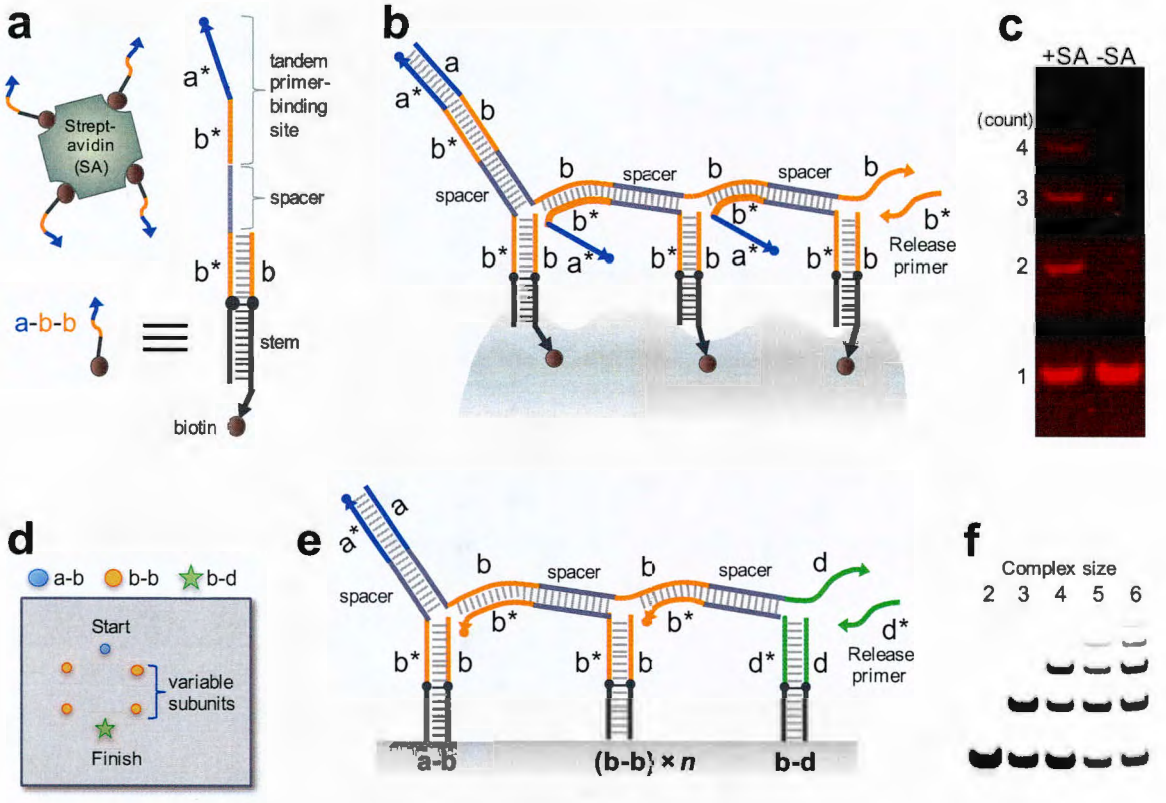


Figure 2

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

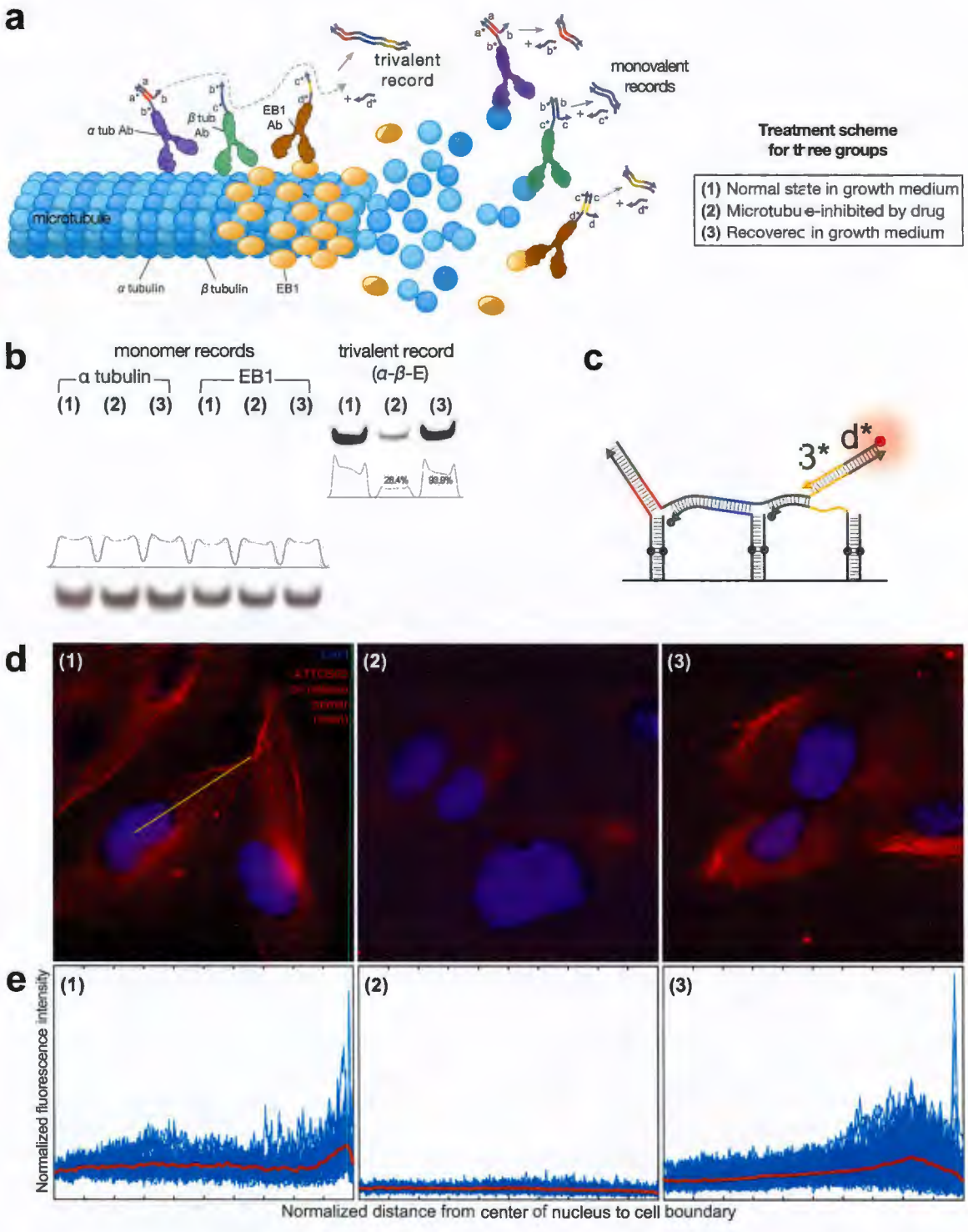


Figure 3

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

Appendix 2

A DNA nanoscope that identifies and precisely localizes over a hundred unique molecular features with nanometer accuracy

Nikhil Gopalkrishnan[1,2], Sukanya Punthambaker[1,3], Thomas E. Schaus[1], George M. Church[1,3] and Peng Yin[1,2,*].

[1]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA, [2]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA, [3]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA.

*Correspondence to: py@hms.harvard.edu

Abstract: Techniques that can both spatially map out molecular features and discriminate many targets would be highly valued for their utility in studying fundamental nanoscale processes. In spite of decades of development, no current technique can achieve both nanoscale resolution and discriminate hundreds of targets. Here, we report the development of a novel bottom-up technology that: (a) labels a sample with DNA barcodes, (b) measures pairwise-distances between labeled sites and writes them into DNA molecules, (c) reads the pairwise-distances by sequencing and (d) robustly integrates this noisy information to reveal the geometry of the underlying sample. We demonstrate our technology on DNA origami, which are complex synthetic nanostructures. We both spatially localized and uniquely identified over a hundred densely packed unique elements, some spaced just 6 nm apart, with an average spatial localization accuracy (RMS deviation) of ~2 nm. The bottom-up, sequencing-enabled mechanism of the DNA nanoscope is fundamentally different from top- down imaging, and hence offers unique advantages in precision, throughput and accessibility.

Introduction

The study of complex materials with nanoscale features benefits from forming an image of it, if possible, with increasingly sophisticated instruments which provide molecular-level detail for further understanding or validation. Comprehensive visualization can be challenging for two reasons – size and molecular diversity. The finest molecular details can only be resolved with nanoscale localization. At the same time there is a tremendous diversity of molecular targets, necessitating the ability to identify and discriminate between these functional components (Fig. 1a). We propose a novel technique, which we term a DNA nanoscope (Fig. 1b), that tags targets with synthetic DNA barcodes, measures distances between many target pairs using biochemical DNA reactions, and then reconstructs a detailed map of the underlying geometry that uniquely identifies every target. We developed and demonstrated the capabilities of the DNA nanoscope technique on 'DNA origami'[1], which are complex synthetic nanostructures composed of hundreds of unique components.

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

Like the DNA nanoscope, long-established and widely used class-average tools like X-ray crystallography and cryoEM also exploit sample periodicity or particle homogeneity to obtain nanometer or even angstrom resolution class-average reconstructions. However, they produce monochromatic images and can only discriminate molecular targets when they are resolved to near atomic precision, unachievable for many samples. At the other end of the spectrum, biochemical techniques like Hi-C can discriminate millions of DNA targets on chromosomes by sequencing them, however the contact densities currently obtainable from single nuclei Hi-C experiments preclude synthesizing this information into a structural model of the chromosome, while geometric models synthesized using data from ensemble Hi-C experiments have at best a local resolution of 5 kilobase pairs[6,7].

In this work, we pooled together information from many identical (up to manufacturing imperfections) copies of DNA origami to construct a class-average image. DNA origami structures have previously been characterized by AFMs[1], EMs[2,3] and super-resolution microscopes[4,5]. However, unlike the DNA nanoscope, none of these techniques can uniquely identify the over one hundred sequence-specific features that make up a typical DNA origami. In the final section we discuss how the DNA nanoscope can be extended from an ensemble technique to a general technique that does not rely on particle homogeneity.

Results

Encoding distances in DNA molecules: At the heart of the DNA nanoscope is a molecular '*ruler*' mechanism that measures the distance between a pair . This fledgling, 'imaging-by-sequencing' field had two main experimental results. Our previous 'auto-cycling proximity recording' (APR)[8] effort demonstrated seven-point reconstructions (spaced ~30 nm apart) from simple, binary proximity data. The subsequent 'DNA microscope', a reaction-diffusion scheme, demonstrated thousands of single particle localizations but only ~10 μm resolution[9]. In contrast to these previous attempts, our DNA nanoscope leverages the particle homogeneity of DNA origami to produce a nanoscale-resolution spatial map of a hundred or more points by making thousands of independent, pairwise distance measurements. This fine resolution is a direct consequence of two novel features of our molecular mechanism. First, our pairwise measurements report distance with high precision, as against previous efforts, which lacked precise distance reporting. Second, the DNA nanoscope measures and reports distances in the 10 nm to 100 nm range that is most relevant to molecular assemblies. This allowed us to resolve large gaps between components, situate otherwise disconnected clusters of points, and build a nanoscale precise global spatial map of the underlying geometry. Simulations showed (see Supplementary Fig. 1) that recording proximities with long reach but little distance precision resulted in maps in which points collapsed into unresolved clumps. Conversely, short reach but higher precision in distance measurements could resolve local geometry but not correctly situate distant points. However, when reach exceeded all gaps between adjacent points and the precision of distance measurement neared ~10%, reconstructions became surprisingly complete and accurate.has of DNA-labeled targets and encodes it in a double stranded DNA molecule, which we call a '*distance record*'. The length of the distance record, in base pairs, directly corresponds to the physical distance measured. The molecular ruler mechanism consists of three stages – growth (Fig. 2a and Supplementary Fig. 2), connection and release (Fig. 2b and Supplementary Fig. 2). Given targets tagged with DNA handles, recording primers are introduced which bind to the handles via hybridization.

Recording primers come in two complementary flavors, with either the sequence domain 'a' or the domain 'a*' (the reverse complement of 'a') at their 3' ends. The recording primers, with the aid of a corresponding extension hairpin, a strand displacing DNA polymerase (Bsm large fragment) and dNTPs, undergo

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

polymerase exchange reactions (PER)[13] which repeatedly add the single-stranded domain sequence 'a' or 'a*' to their 3' end (Fig. 2a and Fig. 2b(1)). Once complementary extended recording primers are long enough, they hybridize to each other via the domains 'a' and 'a*' (Fig. 2b(2)). At this point, again with the aid of a polymerase and dNTPs, the extended recording primers use each other as templates and polymerize to produce a double stranded DNA molecule that is displaced into solution (Fig. 2b(3)). The DNA molecule is a distance record, with the repeat domain 'a a ... a' sandwiched between handle domains. The length of the repeat domain directly corresponds to the physical distance being measured. The sequence of the handle domain can serve as a DNA barcode to uniquely encode the molecular identity of the target. This process of growth, connection and release is isothermal and autonomous.

We wish to stress that each record molecule is a distance measurement, unlike the DNA microscope[9] that encodes distance non-linearly in the *number* of identical proximity-dependent records produced. Of course, not every measurement of the same distance produces a distance record of the same length, because the growing single-stranded recording primers are entropic springs and their growth process is stochastic. However, repeated, independent measurements of the same distance can be aggregated to ultimately provide a ~1 nm distance measurement accuracy. In this work, our targets were positions on a DNA origami[1], a well- characterized nanoscale breadboard. We made repeated measurements by having many identical (up to manufacturing imperfections) copies of DNA origami in the same reaction pot.

Calibration of the molecular ruler: The molecular ruler mechanism produces distance records, whose length, in base pairs, must be related to the physical distance, in nanometers. We performed this translation by means of a calibration experiment. We placed molecular targets at known distances and performed DNA nanoscope recordings, which produced distance records that were then, in aggregate, related to the known distance, yielding a calibration function.

The calibration experiment was performed on a DNA origami adhered to a flat surface. The origami is composed of planar, parallel DNA double helices. The helices are held together by *staple* strands that cross over between neighboring helices at regular intervals along the helical axis. Each staple strand is uniquely addressable by way of its sequence and can be extended into handle domains that serve as handles for recruiting recording primers. We fixed the position of one of the targets near one end of one of the helices of the origami and offset the other target at regular intervals along this helix (Fig. 3a). The handle extended away from the surface and recording primers were bound to it by hybridization. The distance between targets can be calculated simply as the rise per base pair (= 0.34 nm) times the number of base pairs that separate them. The experimental workflow for a calibration experiment was as follows.

The DNA origami was deposited (Fig. 3a(1)) onto a charged, atomically flat mica surface to minimize any flexibility due to thermal motion and reduce variability from one molecular measurement to the next. Reaction components (extension hairpins, strand displacing polymerase and dNTPs) were flown in and a DNA nanoscope recording was performed (Fig. 3a(2)). The distance records produced by this recording were collected, amplified by PCR and characterized by gel electrophoresis (Fig. 3a(3)). The distribution of lengths obtained reflects various independent measurements of the same distance. Our experiments showed that the distance records were skew-normal distributed (Fig. 3b). The greater the distance being measured, the longer, on average, were the distance records produced. The spread of the distribution also widened with increasing distance. We chose the location of the peak of the distribution, which is the most frequently produced distance record, as the representative for the distribution. A low dimensional, monotonically increasing function ($a\sqrt{x}+b$) was fit to the peak data to generate a calibration function (Fig. 3c) that translates distance records into distance measurements. The lengths of the distance records, characterized by next-generation nanopore sequencing, were in excellent agreement with gel measurements (Supplementary Fig. 4).

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

Full-color, molecular scale reconstruction of complex patterns

Armed with a molecular ruler mechanism to produce distance records and a calibration function to convert those records into physical distances, we applied the DNA nanoscope to reconstruct patterns on a DNA origami surface. We uniquely labeled each target feature of the pattern using DNA barcodes; recorded pairwise distances between labeled targets and reconstructed the pattern with molecular resolution.

Pattern design: A DNA origami surface can be abstracted as a hexagonal grid, where each grid point corresponds to the 3' end of a staple strand. A pattern is simply some subset of points chosen from this grid. Fig. 4a(1) shows a Smiley face pattern. All grid points have unique identities associated with them, furnished by the specific DNA sequence of the staple strand at that location.

Tagging: The DNA origami pattern is first prepared for recording by a tagging strategy that associates a barcode sequence with a staple strand. This barcode sequence, synthesized as a 3' appendage on the corresponding staple strand, is used as a 'handle' to specifically recruit, via hybridization, a barcoded recording primer for a ruler measurement. We did not attempt to tag every feature of the pattern in every copy of the DNA origami. Instead we pursued a sparse tagging strategy, where every feature was randomly labeled with some probability and otherwise left unlabeled (Fig. 4a(2)). See Supplementary Materials and Methods for details on how this was achieved. In aggregate, across all the copies of the few hundred thousand DNA origami that were part of an experiment, we expect each feature of the pattern, on average, was tagged thousands of times.

Recording: Once tagged, the DNA origami were deposited on a mica surface, as before, to reduce thermal molecular fluctuations and variations between origami copies. A molecular ruler recording was performed and distance records generated (Fig. 4a(3)). The distance records contained DNA barcodes at either end, corresponding to the underlying targets from which the measurement was produced.

Reading distance records: Finally, both the lengths and barcode sequences of the distance records were read with next generation sequencing. Each distance record was parsed to identify its barcode sequences and then assigned to the target pair from which it was likely generated. The length distribution of distance records for a target pair reflects several independent

measurements of the distance between them. The distribution was smoothed and the location of the most prominent peak extracted. The calibration function was used to translate this peak location into a physical distance measurement.

Inferring geometry from distance data: The question of integrating noisy, pairwise distance measurements into an embedding in Euclidean space, referred to variously as the distance geometry problem[14], global positioning problem[15], localization problem[16] etc., is well studied and has applications in sensor network localization, manifold learning and reconstruction of protein conformation from NMR data. Noisy distance measurements tend to end up in conflict with each other. The problem of producing an accurate embedding is thus a problem of balancing conflicting measurements.

The accuracy of a distance measurement between any two points depends on the number of corresponding distance records read. In fact, we found that the height of the most prominent peak, in units of number of reads, serves as an effective proxy for accuracy. In a calibration experiment we aggregated many thousands (sequencing) to millions (gel electrophoresis) of distance records from a single target pair, allowing us to precisely and accurately pinpoint the peak of the distribution. In contrast, a typical pattern reconstruction experiment only aggregates tens to hundreds of distance records per target pair, increasing uncertainty and producing less accurate distance measurements. Additionally, the multiplexed recording, amplification and

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

sequencing process results in spurious reads, i.e. reads that likely come from unwanted side reactions. In cases where very few distance records are read from a target pair, these spurious reads exacerbate uncertainty and lead to highly inaccurate measurements (Supplementary Fig. 5). However, even with inaccurate data we managed to produce surprisingly accurate reconstructions by identifying less accurate measurements using peak heights and discounting them using weights.

Briefly, if the height of the most prominent peak is below a threshold parameter, the measurement is assigned a weight of 0. As the height exceeds the threshold, the assigned weight asymptotically approaches 1. A smaller weight reduces the influence of the corresponding measurement, allowing us to resolve conflicts in favor of more accurate measurements.

The value of the threshold parameter controls the relative influence of the measurements and changes the produced embedding. A threshold that is too low risks allowing too many inaccurate measurements to influence the embedding while a too high threshold risks discounting too many accurate measurements. The appropriate balance for the threshold parameter is auto-set by trying a series of values, each of which produces an embedding. Each embedding is evaluated for internal consistency, i.e. how well the embedding agrees with the measured distances. The threshold that produces the most internally consistent embedding is chosen as the optimal. Note that this auto-set procedure is without reference to any knowledge of the answer (see Supplementary Note 2C for details on how weights are calculated and Supplementary Note 2D for details on how the threshold parameter is auto- set).

In formal terms, we modeled the question as a global nonlinear optimization problem in two-dimensional Euclidean space. The objective function, which we seek to minimize, is defined as the weighted mean-squared-error between the measured and the embedded Euclidean distance. Simply attempting to minimize the objective function by starting from an arbitrary initial embedding was not robust. The optimization solution space is high-dimensional ($2n$ where n is the number of points in the pattern) and highly non-convex, which resulted in the algorithm being trapped at local minima or saddle points. One solution to this issue would be to do repeated random initializations and pick the best instance. However, this is a computationally expensive approach.

Instead, we solved this optimization problem in three stages. First, points with very few (less than three) reliable (i.e. zero weight) measurements were dropped from the reconstruction, along with all their associated measurements (see Supplementary Note 2A). In the second stage, we used a robust facial reduction algorithm[17] to obtain an initial solution that gave equal weight to all remaining measurements (see Supplementary Note 2B). Weights were then introduced and this initial solution was refined using a quasi-Newton algorithm to find the minimum of the objective function and thus obtain the final reconstructed pattern (see Supplementary Note 2C). The obtained embedding of the points was superimposed on the designed pattern using the Kabsch rigid transformation, which minimizes the root-mean-square deviation, a measure of the average distance between two paired sets of points. We emphasize that the patterns were reconstructed only using information from the pairwise distance records. No *a priori* knowledge of the geometry of the points was used to arrive at the final answer.

Molecular resolution reconstructions: We successfully applied the DNA nanoscope technique to nine different patterns (Fig. 4) and obtained molecular resolution reconstructions. The root-mean-square deviation (RMSD) was used to quantify the average error between the designed and the reconstructed pattern. The RMSDs for the various patterns range from 1.4 nm to 2.6 nm. The points in the most densely packed patterns (Fig. 4b, Rectangle, Chevron, Donut and Pacman) are merely 6 nm apart and yet were clearly spatially resolved in the respective reconstructions. We successfully resolved negative space (Fig. 4a Smiley; Fig. 4b Donut, Frame, Fractal and Pacman), clusters of segregated points (Fig. 4b Frame, Wyss and Pacman) and sparse patterns (Fig. 4b Frame, DNA and Wyss). The variety of patterns

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

reconstructed demonstrates the robustness of our approach. The biggest patterns were approximately 100 nm wide and 50 nm tall. The highest number of points localized was 135, in the Pacman pattern. Apart from spatially localizing the various points of a dense nanoscale pattern, the DNA nanoscope also uniquely distinguishes them by means of their barcode sequence, something that has proven unfeasible for microscopy techniques, which suffer from low- multiplexing capabilities.

Full-color reconstructions: DNA origami has found wide use as a nanoscale breadboard and been decorated with receptor ligands[18], gold nanoparticles[19–21], quantum dots[22], fluorescent dyes[23–25] and carbon nanotubes[26].

The attachment is usually mediated by using an auxiliary sequence tag. These auxiliary tags are independent of, and in addition to, the barcode tags associated with staple strands. Auxiliary tags allow us a programmable way to specify the geometry and absolute valency of objects decorated on DNA origami. Auxiliary tags could also be used to encode UMIs (unique molecular identities) that might help distinguish a particular DNA origami from its cohorts. We show that the DNA nanoscope can be used to natively read auxiliary tags in a multiplexed manner, demonstrating its power to discriminate molecular identity. We used a 12 base auxiliary sequence to encode 'color' information and reconstructed two patterns (Fig. 4c and Fig. 4d) that showcase our ability to read many 'colors' in arbitrary conformations.

Robustness of reconstructions: As remarked earlier, the accuracy of our reconstructions exceed what one may naively expect from looking at the quality of the obtained raw data. For example, many distance measurements are significantly inaccurate (Supplementary Fig. 5) but the resulting reconstructions (Fig. 4e) are very accurate. In fact, we can tolerate further deterioration in data quality without suffering a severe loss of reconstruction accuracy. We again reconstructed patterns from the same experimental data as in Fig. 4, this time first deteriorating the data by one of four distinct methods to test how limited data quality is tolerated by the DNA nanoscope (Fig. 5).

First, we reduced the number of records that were used to reconstruct a pattern by randomly sampling fewer and fewer DNA sequencing reads (Fig. 5a), consequently progressively deteriorating the accuracy of distance measurements. We found that 1 million total reads per pattern are sufficient to obtain ~2 nm RMSD for almost every pattern, and further sequencing did not significantly improve accuracy. Almost half the patterns achieve their optimal accuracy with as few as 100,000 reads (e.g. DNA, Smiley, Color wheel, Chevron and Rectangle). An RMSD of ~5 nm was obtained in some cases with as few as 10,000 reads (e.g. DNA, Smiley and Rectangle). A mere 2,000 reads sufficed to reconstruct the 77 point Color wheel with ~5nm RMSD.

We also tested the effects of an uneven deterioration in data by disregarding all distance reads between some fraction of pairs, resulting in the complete absence of respective distance measurements. We found that the random loss of up to 30% of the distance measurements is well tolerated (Fig. 5b) by most patterns and dense, compact patterns could tolerate the loss of almost 50% of measurements.

Third, as opposed to accuracy, we degraded the precision of the distance measurements by binning them. That is, we created equal sized distance bins (For instance [0 nm, 10 nm], [10 nm, 20 nm], and so on) and measurements that lay within each bin were approximated to the mid-point of that bin (5 nm, 15 nmandsoon).Notethatabinsizeof$l$leadstoan average perturbation of l / 4 in the distance measurements, assuming a uniform distribution of distances in a bin. A larger bin size corresponds to lower precision. We found that precision deteriorations corresponding to bin sizes of up to ~25 nm were well tolerated by the DNA nanoscope (Fig. 5c). In the limit of large bin sizes, we are effectively simulating a proximity-only measurement. Reconstructions fail to be accurate in these cases, demonstrating that in general a degree of

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

precision, i.e. measuring distances and not just recording binary proximity, is critical for accurate reconstructions.

Finally, we confirmed our hypothesis that recording short-range distances that only span immediate neighbors does not produce accurate reconstructions. We simulated this limited 'reach' by discarding all measurements greater than a certain maximum and attempted to reconstruct patterns. We found that when reach extends beyond immediate neighbors (often sufficient to span the larger gaps), reconstruction accuracy improves significantly (Fig. 5d). This suggests that while individual distance measurements may mislead, there is "wisdom in the crowd".

Discussion

We have devised a DNA nanoscope, a tool that records molecular identities and spatial organization in DNA molecules with nanoscale localization accuracy. This DNA nanoscope was used to record dense, nanoscale patterns on homogenous DNA origami particles containing over a 100 unique features. Features spaced just 6 nm apart were clearly resolved with an average spatial localization accuracy of ~2 nm. Each feature was uniquely identified. This combination of spatial resolution and unique molecular identification is unprecedented even for homogenous particles, and has not been achieved by any other technology.

Bottom-up 'imaging by sequencing' technologies, like our DNA nanoscope, stand in contrast to top-down microscopy methods, and confer unique performance and operational advantages. The molecular recording processes that generate spatial data are isotropic and hence we expect that our 3D spatial resolution will match our demonstrated ~5 nm 2D spatial resolution as long as appropriate calibration standards are used. This is in contrast to microscopy methods that have worse 3D resolution in comparison to their 2D resolution. The recording 'instruments' of the DNA nanoscope are a swarm of molecules, diffusing throughout and inspecting a large population of molecular targets in parallel. This eliminates the need to correct any sample drift with respect to the instrument, which imposes practical and fundamental limits on the resolution of microscopy techniques. This parallelism also means that the recording throughput of a large sample is similar to that of smaller samples. The throughput is limited only by our ability to quickly sequence the records. Sequencing throughput is constantly improving and has seen Moore's law like improvements in the past few years.

Apart from these performance advantages, the DNA nanoscope protocol has several operational advantages. First, there is no requirement that the sample be accessible to electromagnetic radiation, only to tiny diffusing DNA molecules that can likely penetrate to otherwise inaccessible locations. Second, the recording interactions with the sample are via gentle biochemical reactions (DNA hybridization and polymerization) in contrast to high-energy lasers, electron beams or physical probes used in super-resolution microscopy, electron microscopy and scanning probe microscopy respectively. Third, the sample does not need any special preparation, like adhering it to a surface, or freezing it in vitreous ice, that hold it immobile with respect to macro-scale recording instruments. The recording process is setup simply like a PCR reaction, except without any temperature cycling. Fourth, no capital-intensive, complex and hard to maintain instruments that are periodically rendered obsolete need to be purchased. A $1000 start-up kit available from a commercial source was used in this work. The per assay cost is currently high, costing about $500 per structure mapped in this work, but is seeing rapid drops in price as the technology continues to mature.

The bottom-up 'imaging by sequencing' field is nascent, and many challenges and opportunities remain. In this work, we exploited the homogeneity of DNA origami to reconstruct class average structures. The technique can potentially be extended to acquire images of the structure of single particles. Currently, our ruler recording is pairwise destructive and only one copy of a distance record can be generated from each

labeled target. This precludes single particle reconstructions, as the resulting disjoint pairwise distances cannot be integrated into a spatial map. One solution is to combine pairwise non-destructive recording, as described in our previous 'APR'[8] work, with the 'molecular ruler' mechanism demonstrated here.

As we scale down to single particle reconstructions we scale up in the number of molecular features that we must resolve. In class average experiments, distinct physical copies of molecular targets are superimposed and treated as one target. In contrast, in single particle experiments, each physical copy will have to be treated separately. Thus, an experiment may have millions of unique targets. We argue that the DNA nanoscope technique will scale up to these numbers. Consider the case of a typical DNA origami experiment, with 50,000 DNA origami structures, each consisting of 50 points. These 2.5 million (50,000 times 50) targets can be labeled with unique DNA barcodes (there are a possible ~1 billion DNA sequences of length 20). Ruler recording reactions occur asynchronously and in parallel. Each ruler reaction at least a minute to produce a distance record[8]. Thus, a DNA origami can produce, on average, a few thousand distance records in a matter of hours. We have shown that 10,000 distance records per DNA origami proved sufficient to reconstruct them with sub-5nm accuracy. Improvements in the ruler mechanism that narrow the distribution of record lengths produced for each distance will further reduce the sequencing requirements. The total number of records that would need to be sequenced would be on the order of 500 million, already in reach of short-read sequencing technology and only an order of magnitude away from what long-read nanopore sequencing can currently achieve. The sequencing library could also be split, with shorter records sequenced on short-read high volume sequencers and the long-read nanopore devices focused on the longer reads.

We predict that the DNA nanoscope and related 'imaging by sequencing' techniques will gain widespread adoption over the next few years and drive fundamental nanoscale discoveries.

Materials and Methods

A brief summary of the methods is provided here. Additional details may be found in the Supplementary Materials and Methods section.

DNA origami manufacture and purification: The scaffold strand (M13mp18 single stranded DNA, 5 nM final concentration) was combined with: (i) all 216 'blunt' staple oligos (50 nM final concentration of each oligo, see Supplementary Table 2 for sequences), (ii) the appropriate subset (depending of the pattern being tagged, see Supplementary Fig. 7, 8 and 9 and Supplementary Table 3) of barcoded 'handle' staple oligos (5 nM final concentration of each oligo) and (iii) corresponding appropriate subset of barcoded primers of type a and a* (5 nM final concentration of each oligo, see Supplementary Table 4 for sequences) in 1X TE Mg buffer (pH 7.4, 10 mM Tris-HCl, 0.1 mM EDTA, 10 mM $MgSO_4$). The mixture was then cooled from 90°C to 60°C at the rate of 1 min/°C and then from 60°C to 50°C at the rate 10 min/°C and finally from 50°C to 25°C at the rate of 1 min/°C. Folded origami was stored at 4°C for up to one week prior to purification. DNA origami were purified by agarose gel electrophoresis to eliminate misfolded and aggregated origami as well as to remove excess staple and primer oligos.

DNA nanoscope recording: A thin layer of mica was peeled from a mica sheet using stieky tape and then affixed to a sticky bottomless six-channel slide to assemble fluid-exchange reaction chambers for recording experiments. Purified DNA origami (50 µL at 50 pM) was added to the reaction chamber and allowed to bind to the mica surface for 10 min. The chamber was then washed twice with 50 µL of 1X TE Mg to remove unbound origami. The exposed, unbound mica surface is then passivated with a BSA solution (50 µg/mL) for 5 min and further washed with 1X TE Mg and a magnesium-supplemented 1X Thermopol buffer (20 mM Tris-HCl, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 7 mM $MgSO_4$, 0.1% Triton®-X-100, pH 8.8 @

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

25°C). 50 μL of the recording mix, consisting of 100 nM extension hairpin type 'a', 100 nM extension hairpin type 'a*', 0.08 U/μL Bsm DNA polymerase LF, 100 μM dNTP solution mix, 1X Thermopol buffer and 5 mM $MgSO_4$, is added to the reaction chamber and the slide kept at 37oC for 3 hours. After 3 hours, the supernatant

containing distance records was aspirated and PCR amplified for further characterization. **PAGE characterization:** The length distribution of the distance records for each calibration distance was characterized by running PCR amplified distance records on a denaturing PAGE gel (180 V for 30 min at 50°C). Gel images were analyzed with the Fiji image processing software package.

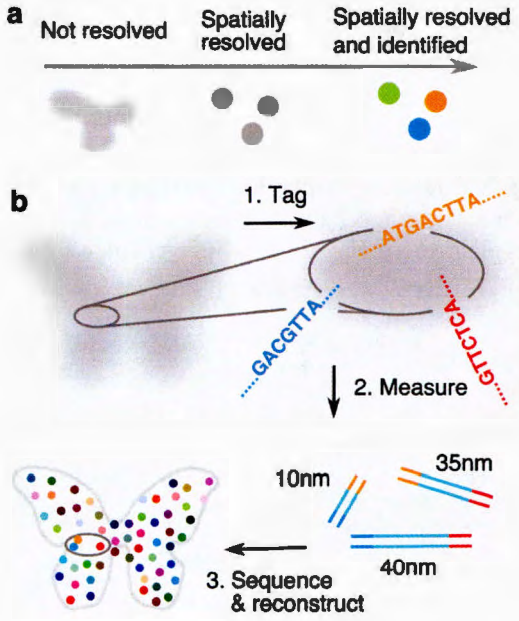
Next-generation sequencing and analysis: PCR amplified distance records were purified by denaturing polyacrylamide gel electrophoresis (see Supplementary Methods for details) to remove short- length spurious distance records. Purified distance records were prepared for next-gen sequencing using the Oxford Nanopore SQK-LSK109 ligation sequencing kit and sequenced to produce 10 to 15 million raw reads. We used Oxford Nanopore's Guppy basecalling software (v3.2.1) to (1) read sequence information from raw sequencing data and then use MATLAB scripts to (2) demultiplex reads from different experiments, (3) extract the lengths of the distance records and assign them to their appropriate target-pair, (4) infer the distance for each target-pair from all assigned distance records, and finally (5) reconstruct the underlying geometry from pairwise distance measurements. The MATLAB scripts can be found at github.com/nikhil314/DNA-Nanoscope.

Acknowledgments: We thank Prof. William M. Shih for useful discussion and comments on this work. **Funding:** This study is supported by grants to P.Y. by the Office of Naval Research (under grants N00014-16-1-2410 and N00014-18-1-2549), the National Institutes of Health (under grants 1R21CA235421-01, 5DP1GM133052-02 and 5R01GM124401-02), the National Science Foundation (under grants CBET-1729397 and MCB-1540214), the Defense Advanced Research Projects Agency (under grant W911NF-17-1-0075), and the Molecular Robotics Initiative at Wyss Institute; **Author contributions:** N.G. conceived and designed the study, performed the experiments, developed the software, collected and analyzed the data, and wrote the manuscript. T.S. conceived the study, analyzed the data, and wrote the manuscript. S.P. performed nanopore sequencing experiments and collected data. G.M.C. provided scientific guidance and contributed to study supervision. P.Y. conceived and supervised the study and wrote the manuscript. All authors edited and approved the manuscript; **Competing interests:** N.G., T.S. and P.Y. have filed a provisional patent covering aspects of this work. P.Y. is co-founder and director of Ultivue Global, NuProbe Inc. and Torus Biosystems; and **Data and materials availability:** The raw sequencing data files are many GB in size and are available on reasonable request. The code for performing the analysis and reconstructions is available on GitHub and a link to it can be found in the Materials and Methods section.

Supplementary Materials available online:

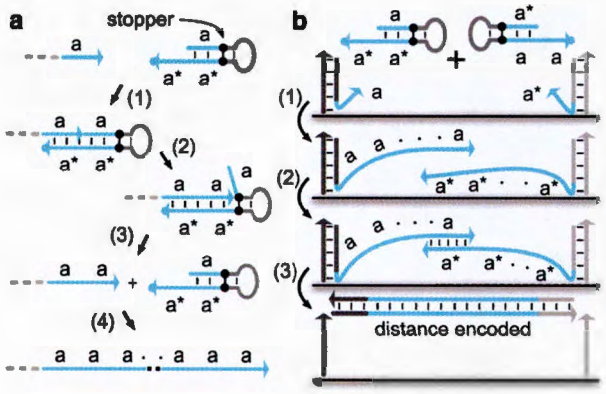
Supplementary Materials and Methods Supplementary Notes 1 and 2 Supplementary Figures 1 to 10 Supplementary Tables 1 to 5 References (27-28)

Fig. 1 The DNA nanoscope 'imaging-by-sequencing' can both distinguish many targets and resolve features at the nanometer scale. **a.** A comprehensive visualization requires that we simultaneously resolve targets spatially and also determine their identity. **b.** Bird's eye view of the DNA nanoscopy process. We tag targets with unique DNA barcodes, measure distances between many target pairs using DNA molecules, read the distances with massively parallel sequencing and integrate them into a molecular resolution spatial map that uniquely discriminates every target.



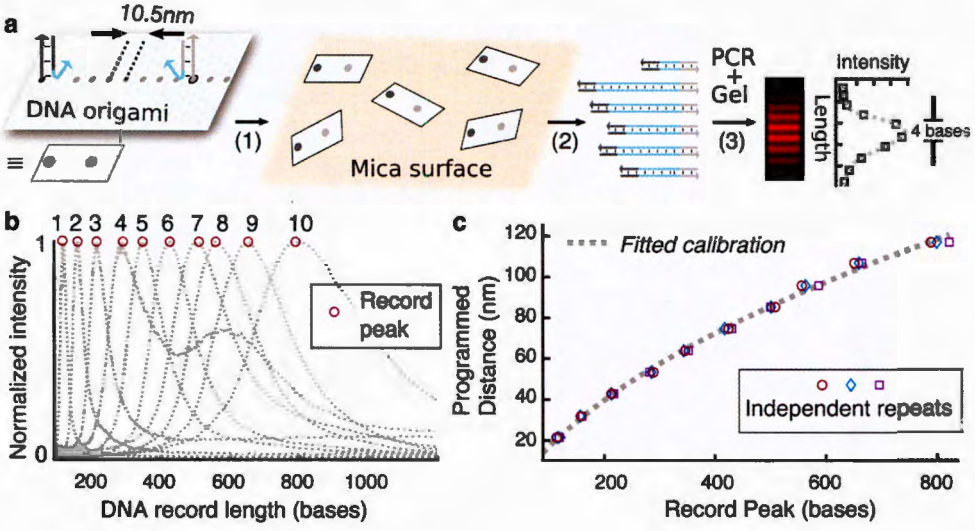
OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

Fig. 2 Molecular ruler mechanism (simplified). **a.** A primer exchange reaction (PER) cascade repeatedly adds the four base sequence domain 'a', as follows. (1) The recording primer hybridizes to a PER hairpin, (2) a strand displacing DNA polymerase (Bsm large fragment) extends the primer into the stem of the hairpin and in the process copies domain 'a'. A 'stopper', a non-canonical base modification on the template that is not recognized by the DNA polymerase, blocks further extension. The polymerase dissociates from the hairpin. (3) The recording primer is only weakly bound to the hairpin and also dissociates. (4) The above sequence of reactions repeat, adding domain 'a' every time. In the same manner, a complementary PER cascade, shown in Supplementary Figure 2, repeatedly adds the four base sequence domain 'a*'. **b.** A double-stranded DNA 'distance record' is generated as follows. Consider two DNA labeled targets with recording primers hybridized to them. (1) The primers take part in PER reaction cascades, as described in part A, adding sequence repeats of 'a' and 'a*' respectively. (2) The extended primers hybridize, (3) copy each other with the aid of the polymerase, are displaced from the targets and released into solution, making a distance record. The molecular ruler mechanism depicted here is a simplification. The full, actual mechanism is depicted in Supplementary Fig. 2. See Supplementary Note 1 for the rationale for our molecular design choices.



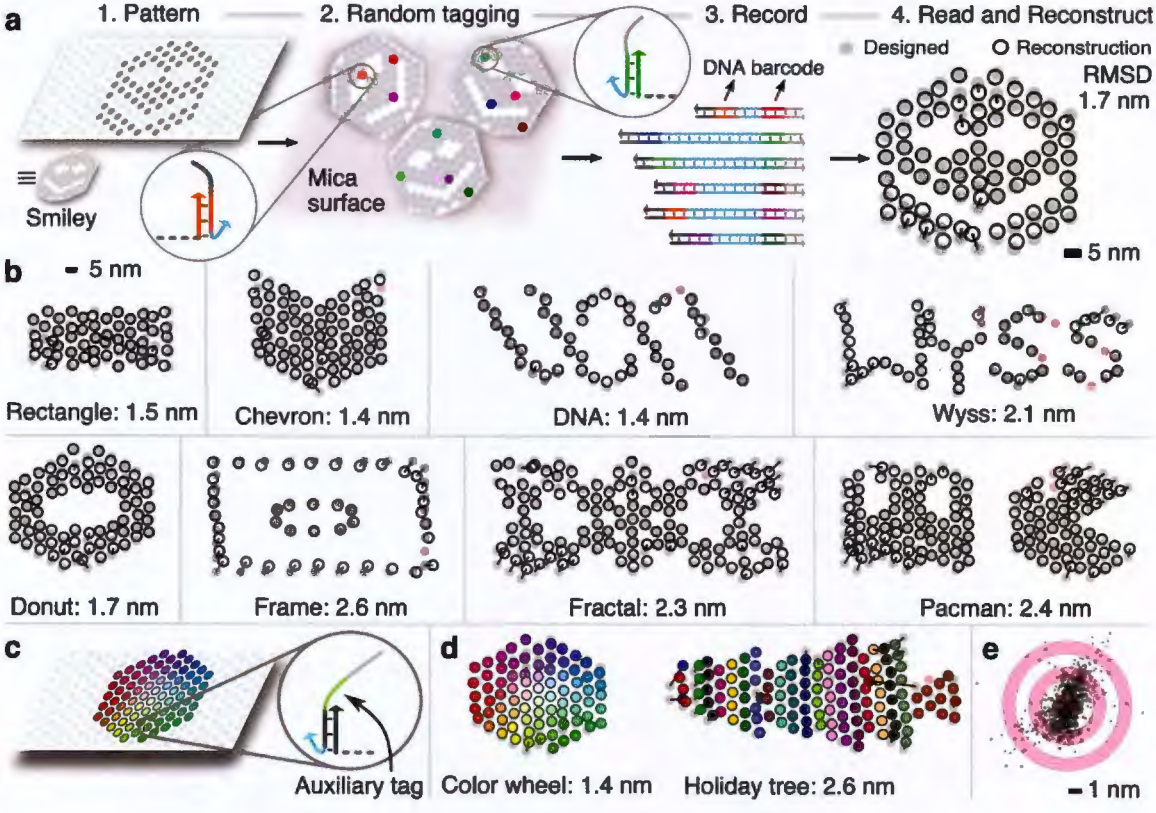
OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

Fig. 3 Molecular ruler calibration. **a.** DNA origami is used as a calibration standard. (1) DNA origami is deposited on a mica surface, (2) ruler recording generates distance records, (3) which are amplified by PCR and characterized by gel electrophoresis, which reveals a skew- normal distribution of record lengths. The discrete bands are 4 bases apart. See Materials and Methods for details on origami design and purification, ruler reaction conditions and PCR and gel protocols. **b.** Gel profiles of record lengths obtained from molecular recordings for various distances between target pairs. The programmed calibration distances are 1 = 21.4 nm, 2 = 32.0 nm, 3 = 42.8 nm, 4 = 53.4 nm, 5 = 63.9 nm, 6 = 74.8 nm, 7 = 85.3 nm, 8 = 95.9 nm, 9 = 106.8 nm and 10 = 117.3 nm. Each profile is normalized to its peak height. Bigger distances produce longer records that are more broadly distributed. The plotted DNA record lengths include primer regions of 32 bases each at either end. **c.** A calibration curve is fit to the peak of the distribution, giving us a function to transform a distance record, in base pairs, into a physical distance, in nanometers. The gel image and corresponding gel profiles for all three independent repeats can found in Supplementary Fig. 3).



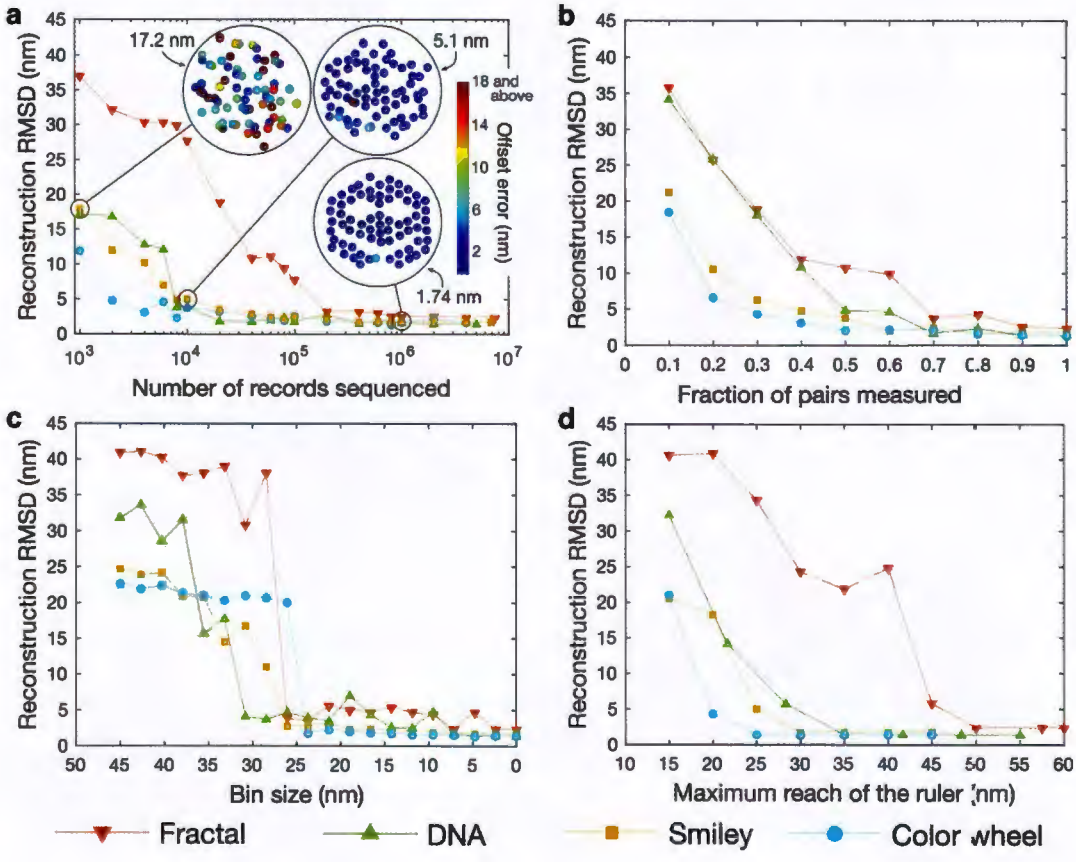
OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

Fig. 4 DNA nanoscope applied to various patterns **a.** (1) A pattern is some subset of positions on the DNA origami chosen from the underlying hexagonal grid. (2) A random subset of points is tagged with barcoded primers. While positions within each origami are only sparsely tagged, in aggregate each position is tagged many times over. (3) The identity of targets as well as the distance between targets is encoded inside barcoded distance records. (4) Distance records are read with next-generation sequencing to obtain length and barcode information, which is used to infer distances between points. An algorithm integrates pairwise distance measurements into a nanoscale precise map by embedding the points in a Euclidean plane. The reconstruction (hollow circles) is overlaid on the designed pattern (gray solid circles) for comparing the accuracy of the reconstruction. **b.** Many different patterns reconstructed with high accuracy. Each pattern is drawn to the same scale (scale bar = 5 nm). The numbers below the pattern are the RMS deviation between the designed and reconstructed pattern. Points missing from the reconstruction are indicated with red solid circles as opposed to gray solid circles. **c.** We encoded 'color' in auxiliary sequence tags that were then read out with the DNA nanoscope. **d.** Color wheel pattern with 77 distinct colors. Each auxiliary tag is unique. Holiday tree with 21 distinct colors. Each separate column of the pattern is a distinct auxiliary sequence, while points within the same column share the same sequence. All 13 points that make up the trunk share the same auxiliary sequence. **e.** An aggregate view of the accuracy of all the reconstructions from **b** and **d**. Each dot corresponds to the offset error vector between the reconstructed and the designed point. Each offset vector is translated to the center of the bulls-eye, whose each ring is 1 nm wide.



OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

Fig. 5 Reconstruction accuracy with deteriorating data quality for four representative patterns (Fractal, DNA, Smiley and Color wheel). **A.** We reconstructed patterns by successively sampling fewer and fewer sequencing reads. Even 10,000 sequence reads are sometimes sufficient to obtain ~5nm or better accuracy. Each plotted RMSD is an average of 10 independent samples. The inset shows example reconstructions. Each point is shaded by its error, which is defined as its offset distance from its designed position. **B.** The loss of some fraction of pairwise distance measurements, chosen at random from all possible pairs, is well tolerated by the DNA nanoscope. **C.** Distance measurements are binned to reduce precision. Some precision is necessary to reconstruct patterns with high accuracy, but the accuracy of the reconstruction does not significantly deteriorate with some loss of precision. On way of understanding the quantitative effect of bin size is to note that a bin size of l introduces an average error of $l / 4$ in a measurement, assuming a uniform distribution of distances within a bin. Thus, a bin size of 20nm would introduce an average error of 5nm in the measurements. **D.** All measurements between points farther apart than a maximum reach are discarded to demonstrate the effect of limited ruler reach. The closest spaced points in our patterns are 6 nm apart. We observe that a maximum reach limited to immediate neighbors fails to produce high-quality reconstructions. When reach extends beyond immediate neighbors, construction quality significantly improves. A reach extending to span the diameter of the pattern did not significantly improve reconstruction accuracy.



OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

References

1. Rothemund, P. W. K. Folding DNA to create nanoscale shapes and patterns. *Nature* **440**, 297–302 (2006).
2. Douglas, S. M. *et al.* Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature* **459**, 414–418 (2009).
3. Bai, X. C., Martin, T. G., Scheres, S. H. W. & Dietz, H. Cryo-EM structure of a 3D DNA-origami object. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 20012–20017 (2012).
4. Dai, M., Jungmann, R. & Yin, P. Optical imaging of individual biomolecules in densely packed clusters. *Nat. Nanotechnol.* **11**, 798–807 (2016).
5. Iinuma, R. *et al.* Polyhedra self-assembled from DNA tripods and characterized with 3D DNA-PAINT. *Science* **344**, 65–9 (2014).
6. Le Treut, G., Képès, F. & Orland, H. A Polymer Model for the Quantitative Reconstruction of Chromosome Architecture from HiC and GAM Data. *Biophys. J.* **115**, 2286–2294 (2018).
7. Lesne, A., Riposo, J., Roger, P., Cournac, A. & Mozziconacci, J. 3D genome reconstruction from chromosomal contacts. *Nat. Methods* **11**, 1141– 1143 (2014).
8. Schaus, T. E., Woo, S., Xuan, F., Chen, X. & Yin, P. A DNA nanoscope via auto-cycling proximity recording. *Nat. Commun.* **8**, 696 (2017).
9. Weinstein, J. A., Regev, A. & Zhang, F. DNA Microscopy: Optics-free Spatio-genetic Imaging by a Stand-Alone Chemical Reaction. *Cell* **178**, 229-241.e16 (2019).
10. Boulgakov, A. A., Ellington, A. D. & Marcotte, E. M. Bringing Microscopy-By-Sequencing into View. *Trends in Biotechnology* **38**, 154–162 (2020).
11. Boulgakov, A. A., Xiong, E., Bhadra, S., Ellington, A. D. & Marcotte, E. M. From Space to Sequence and Back Again: Iterative DNA Proximity Ligation and its Applications to DNA-Based Imaging. *bioRxiv* 470211 (2018). doi:10.1101/470211
12. Hoffecker, I. T., Yang, Y., Bernardinelli, G., Orponen, P. & Högberg, B. A computational framework for DNA sequencing microscopy. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 19282–19287 (2019).
13. Kishi, J. Y., Schaus, T. E., Gopalkrishnan, N., Xuan, F. & Yin, P. Programmable autonomous synthesis of single-stranded DNA. *Nat. Chem.* **10**, 155–164 (2018).
14. Souza, M., Lavor, C., Muritiba, A. & Maculan, N. Solving the molecular distance geometry problem with inaccurate distance data. *BMC Bioinformatics* **14**, (2013).
15. Singer, A. A remark on global positioning from local distances. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 9507–9511 (2008).
16. Javanmard, A. & Montanari, A. Localization from Incomplete Noisy Distance Measurements. *Found. Comput. Math.* **13**, 297–345 (2013).
17. Drusvyatskiy, D., Krislock, N., Voronin, Y. L. & Wolkowicz, H. Noisy euclidean distance realization: Robust facial reduction and the pareto frontier. *SIAM J. Optim.* **27**, 2301–2331 (2017).
18. Hawkes, W. *et al.* Probing the nanoscale organisation and multivalency of cell surface receptors: DNA origami nanoarrays for cellular studies with single-molecule control. *Faraday Discuss.* **219**, 203–219 (2019).
19. Liu, W., Halverson, J., Tian, Y., Tkachenko, A. V. & Gang, O. Self-organized architectures from assorted DNA-framed nanoparticles. *Nat. Chem.* **8**, 867–873 (2016).
20. Kuzyk, A. *et al.* DNA-based self-assembly of chiral plasmonic nanostructures with tailored optical response. *Nature* **483**, 311–314 (2012).
21. Acuna, G. P. *et al.* Fluorescence enhancement at 11 docking sites of DNA-directed self-assembled nanoantennas. *Science* **338**, 506–10 (2012).

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

21. Samanta, A., Zhou, Y., Zou, S., Yan, H. & Liu, Y. Fluorescence quenching of quantum dots by gold nanoparticles: A potential long range spectroscopic ruler. *Nano Lett.* **14**, 5052–5057 (2014).
22. Gopinath, A., Miyazono, E., Faraon, A. & Rothemund, P. W. K. Engineering and mapping nanocavity emission via precision placement of DNA origami. *Nature* **535**, 401–405 (2016).
23. Lin, C. *et al.* Submicrometre geometrically encoded fluorescent barcodes self-assembled from DNA. *Nat. Chem.* **4**, 832–839 (2012).
24. Hemmig, E. A. *et al.* Programming Light- Harvesting Efficiency Using DNA Origami. *Nano Lett.* **16**, 2369–2374 (2016).
25. Maune, H. T. *et al.* Self-assembly of carbon nanotubes into two-dimensional gecmetries using DNA origami templates. *Nat. Nanotechnol.* **5**, 61– 66 (2010).

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

Appendix 3

Light-controlled DNA barcoding

Summary

Here, we present a new approach for **light-directed DNA barcoding**. Our method consists of building the DNA barcodes on surfaces by light-controlled multi-strand concatemerization. This label ng approach fuels a new workflow that can scale the number of features that can be uniquely addressed with a method of combinatorial barcoding (Figure 1).

Technical Description and Current Progress

Using the photocrosslinking chemistry that has already been validated in the Action-PAINT workflow (Liu et al. Nature Chemistry, 2009), we can utilize the spatial control of light with Digital M cromirror Devices (DMD) to illuminate arbitrary regions of interest (Figure 2). This has the immediate advantage of parallelizing the labeling workflow by independently illuminating and labeling multiple regions at once. Individual feature sizes can be reduced down to the diffraction limit of light (~250 nm) and has currently been validated for micron sized labeling features on a dot array (Figure 3).

Multi-round barcoding will eventually require the use of exchanging DNA barcodes. Typically, a buffer exchange process is performed to wash out unused barcodes from one cycle and introduce a second barcode strand for a new round of labeling. To validate this in a proof-of-concept demonstration, we performed a three round labeling cycle of three parts of a penrose triangle with different fluorescently labeled DNA barcode strands. Once the cyclical labeling workflow was finished, the final image was reconstructed and all fluorescent channels overlaid to synthesize the complete penrose triangle image (Figure 4).

Multi-round combinatorial barcoding will scale exponentially with the number of labeling cycles that will be performed. The general strategy is detailed in Figure 5, where a trit based barcode system (0, 1, 2) was combinatorially labeled over 7 cycles. Such a scheme can provide a feature space of over 2000 unique

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

features. Figure 5 illustrates how each labeling cycle can leverage the independent illumination profile of a DMD to parallelize labeling across a hundred individual features, each labeling cycle will then concatenate more barcode sequences such that by the final cycle, every labeled spot would have a unique barcode sequence that can be reconstructed through sequencing.

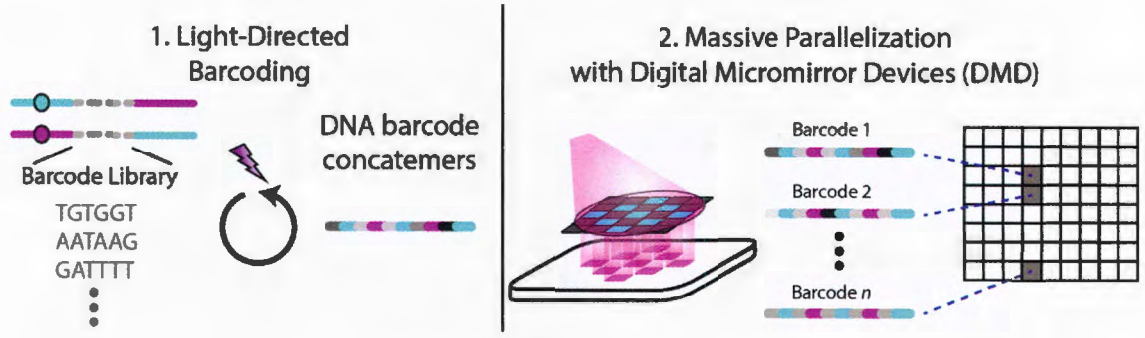


Figure 1: Strategy of multi-strand concatemerization of light-directed barcoding.

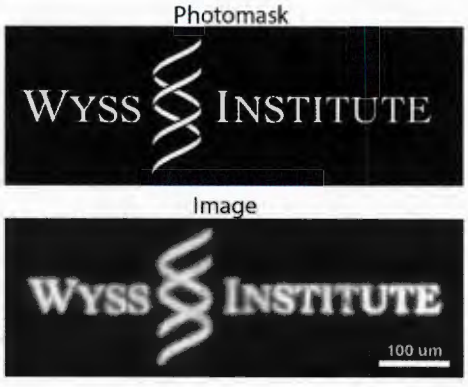


Figure 2: Spatially directed labeling using custom photomasks. A photomask was uploaded to the DMD device and a subsequent image of the fluorescent labeling strands was taken to confirm labeling.

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

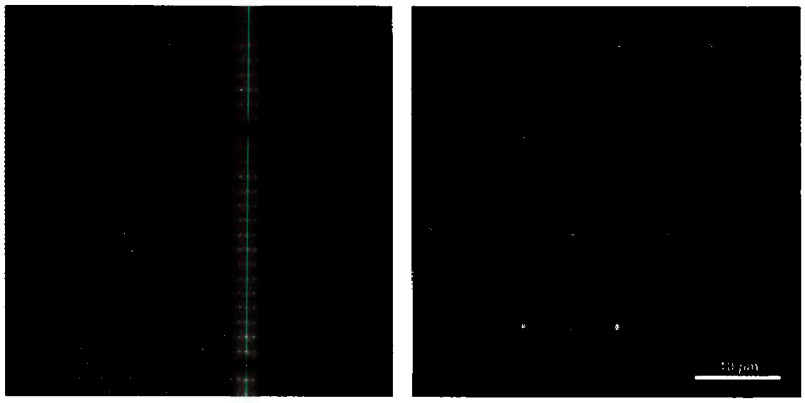


Figure 3: Dot Array labeling on surface demonstrating micron scale feature size.

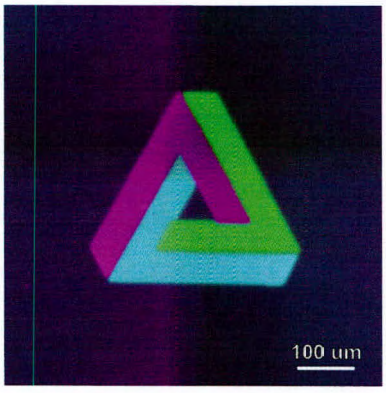


Figure 4: Multi-cycle orthogonal labeling. Three photomasks were labeled in succession to reconstruct an image illustrating a Penrose Triangle.

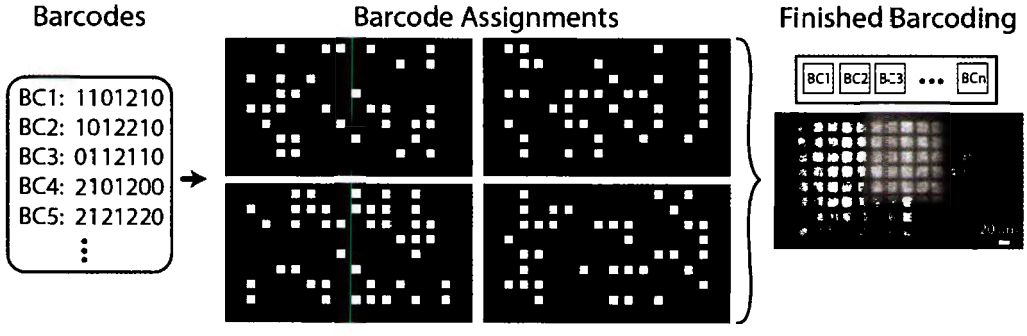


Figure 5: Parallelized barcoding workflow, whereby each feature can be uniquely addressed with a multi-strand barcode sequence.

OMB Control Number: 0704-0527
OMB Expiration Date: 06/30/2025

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
12-12-2022	FINAL	09/15/2018-09/14/2022

4. TITLE AND SUBTITLE
DNA-based technologies for reading and writing large-scale molecular patterns with nanoscale-precision

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-18-1-2549

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Peng Yin

5d. PROJECT NUMBER
1000009566

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
President & Fellows of Harvard College
1033 Massachusetts Ave, 5th Fl
Cambridge, MA 02138-5369

8. PERFORMING ORGANIZATION REPORT NUMBER
167992

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
ONR REG BOSTON N62879
495 SUMMER STREET ROOM 627
BOSTON, MA 02210-2109

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
N62879

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for Public Release; distribution is Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In many technologies there is a trend toward simultaneous miniaturization and increased capacity. The amount of data stored on consumer-advertised desktop computer hard drive, for example, has increased exponentially from 0.1GB in 1980 to 1000GB in 2010 while occupying the same volume. In data storage, the benefits scale only linearly with the amount of data fitting in a given volume, but nevertheless there is a strong driving force for development; computer data is currently generated and stored at 10^12 GB (zettabytes) per year worldwide. In other classes of problems, such as biology, benefits appear to increase out of proportion to scaling. For example, the success of animals requires functionality from nanometer-scale enzymes, molecular motors, and selective membrane pores through the meter-scale processes of digestion, nerve conduction, and musculoskeletal mobility. Nature occasionally sees competitive advantages in extending the upper limit of the scale to tens of meters. A fundamental goal in science and engineering, therefore, is to maintain nanometer-scale (molecular) control of matter across length scales of 1 um (1,000-fold range,), 1 mm (1,000,000-fold), or even higher; and a fundamental challenge in biological research is to understand and measure the nanoscale effects of individual biomolecules across large spatial scales.

15. SUBJECT TERMS
DNA nanotechnology, Nanoscale labeling, Nanoscale Imaging, Addressable biosurfaces.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Peng Yin
U	U	U	UU	41	19b. TELEPHONE NUMBER *(Include area code)* 617-432-7731

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.