



BOISE STATE UNIVERSITY

IGEM # 19-002: Nucleic Acid Memory

July 1, 2019 – December 31, 2020 Annual Report

Will Hughes

Tim Andersen

Eric Hayden

Wan Kuang

Will Clay

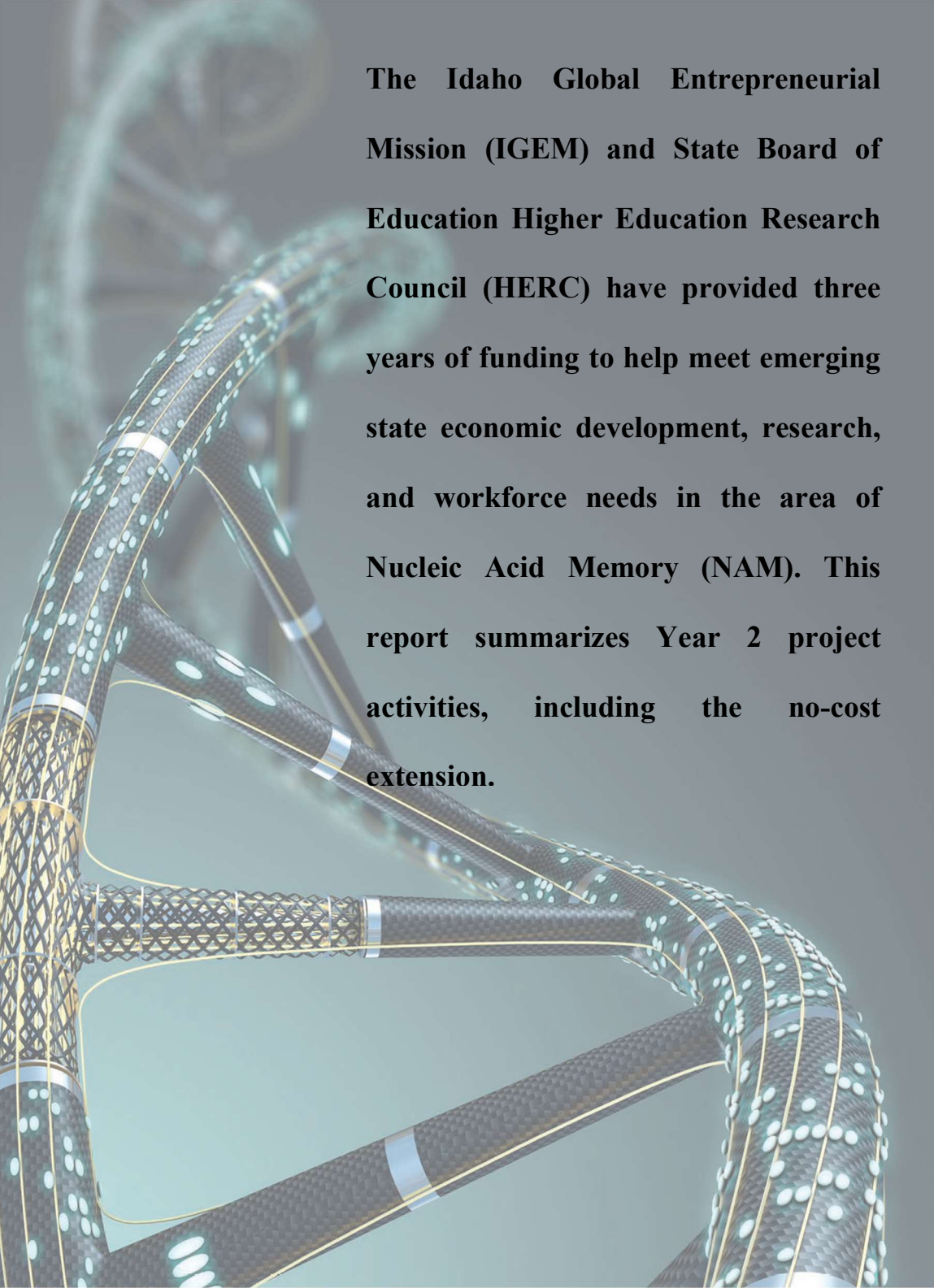
George Dickinson

Luca Piantanida

Mike Tobiason

Chad Watson

I. Project Summary



The Idaho Global Entrepreneurial Mission (IGEM) and State Board of Education Higher Education Research Council (HERC) have provided three years of funding to help meet emerging state economic development, research, and workforce needs in the area of Nucleic Acid Memory (NAM). This report summarizes Year 2 project activities, including the no-cost extension.

II. Project Overview

In 2016, the digital universe produced 16 ZB (1 ZB = 1 trillion GB) of data. In 2025 it will create 163 ZB. These data, once generated, cascade through the information lifecycle — from primary storage media in the form of hard disks and solid-state drives to archival media such as tape. While the semiconductor industry maximizes the density, stability, and energy efficiency of electronic and magnetic memory, both are fast approaching their physical and economic finish lines. As envisioned by the new Semiconductor Synthetic Biology Roadmap, DNA-based massive information storage is a fresh start for memory manufacturing in the United States. According to our study with Micron, Harvard, and the Semiconductor Research Corporation (SRC), DNA has a retention time that ranges from thousands to millions of years, 1 kg of DNA can store the projected digital universe in 2040, and DNA's energy of operation is 100 million times less than current electronic memory. As a result, nucleic acid memory has become a global conversation, a national investment, an industrial opportunity, and a local strength in Idaho.

Our vision is to pioneer a digital data storage paradigm in Idaho by designing, building, and testing accessible, editable, and non-volatile nucleic acid memory (NAM) technologies that are inspired by DNA circuits and made possible by our innovations in DNA nanotechnology. With support from IGEM-HERC, we are creating a Nucleic Acid Memory Institute to meet critical innovation, economic, and workforce development needs in Idaho. To expedite our vision of Idaho becoming a global leader in NAM, five tasks are being addressed over the life of the IGEM-HERC: **Task 1** – Create efficient algorithms for coding information into data strands. Error correction strategies will account for DNA insertions, deletions, and substitutions, as well as screen for biological sequences to ensure that the data has no genetic function. **Task 2** – Create a high-throughput, integrated analytical engine to design and select data strands using quantitative metrics based on an in-house, algorithm. **Task 3** – Create synthetic biological factories for manufacturing DNA scaffolds using rapid design-build-test cycles of genomes. Genome size and structure will be engineered. **Task 4** – Design and fabricate NAM storage platforms using the DNA scaffolds, and validate the functionality of genome scaffolds using atomic force microscopy. **Task 5** – Read arbitrary data files into NAM storage nodes using super-resolution microscopy. Realize sub-nanometer imaging resolution to enable high areal density data storage.

This progress report spans July 1, 2019 to December 31, 2020. Listed below is a summary of our accomplishments during this time period. Because of COVID-19, our team continues to invest into computational work to offset the impact on our ability to perform experimental work.

III. Summary of project accomplishments

The support provided by IGEM-HERC during year 1 of this project provided the infrastructure and team to create the first digital Nucleic Acid Memory (dNAM) proof-of-concept. Building on this foundation — which was described in the July 1, 2018 to June 30, 2019 Annual Report and reported in the July 1, 2019 to January 1, 2020 progress report — we conducted a series of experiments that validated dNAM as a platform for DNA-based information storage. This work led to a manuscript that was submitted to Nature on July 17, 2020. We were notified by the senior editor of Nature on October 26, 2020 that the reviewers did not recommend for publishing in Nature, but instead to consider addressing reviewer comments and submit to Nature Communications, which is an equally respected, widely-read, peer-reviewed journal within the scientific community. We are finalizing the manuscript for resubmission. We also submitted a proposal to the Semiconductor Synthetic Biology for Information Storage and Retrieval (SemiSynBio-II) program, which extends the dNAM platform supported by IGEM-HERC from two dimensions to three dimensions. Our novel approach spatially and temporally reads three-dimensional nucleic acid memory ($3D$ NAM) with sub 5 nm lateral and 1 nm axial resolution. In brief, $3D$ NAM is a synergy between semiconductor technology and synthetic biology—integrating time-correlated super-resolution microscopy and DNA self-assembly to digitally read non-volatile and randomly accessible information in all three dimensions. With information densities above 10 Tbit/cm², read speeds over 56 Tbit/day, and the promise for scalable random access, $3D$ NAM has the potential to be a disruptive memory technology. While we were not selected for funding, in part because we have an active grant through the funding mechanism, the ideas generated and the proposal development process were a galvanizing experience for the team that has led us to explore new techniques and architectures for archival memory storage applications as part of our Year 3 IGEM-HERC deliverables.

The following report details the major work and outcomes supported by IGEM-HERC from July 1, 2019 to December 31, 2020, including updated content within the manuscript in support of the project tasks.

Task 1 – Create improved algorithms for coding information into data strands.

1.1 Encoding/decoding algorithms to create a working prototype of dNAM. dNAM relies on DNA-PAINT to detect individual DNA molecules and is routinely limited by incomplete strand incorporation, defective imager strands, fluorophore bleaching, and background fluorescence. To overcome these challenges, Prof Tim Andersen and graduate student Golam Mortuza created dNAM-specific information encoding and decoding algorithms that combine fountain codes with a custom, bi-level, parity-based, and orientation-invariant error detection scheme (**Fig. 1**). Fountain codes enable transmission of data over noisy channels. They work by dividing a data file into smaller units called droplets and then sending the droplets at random to a receiver. Droplets

can be read in any order and still be decoded to recover the original file, so long as a sufficient number of droplets are sent to ensure that the entire file is received. The custom algorithm they developed encodes each droplet onto a single origami and adds additional bits of information for error correction to ensure that individual droplets will be recovered, in the presence of high noise, from individual origami. Together, the error correction and fountain codes increase the probability that the message is fully recovered while minimizing the number of DNA origami that must be observed. These encoding/decoding algorithms were used to create a working prototype of dNAM — encoding the short message ‘Data is in our DNA!\n’, see Task 5, below.

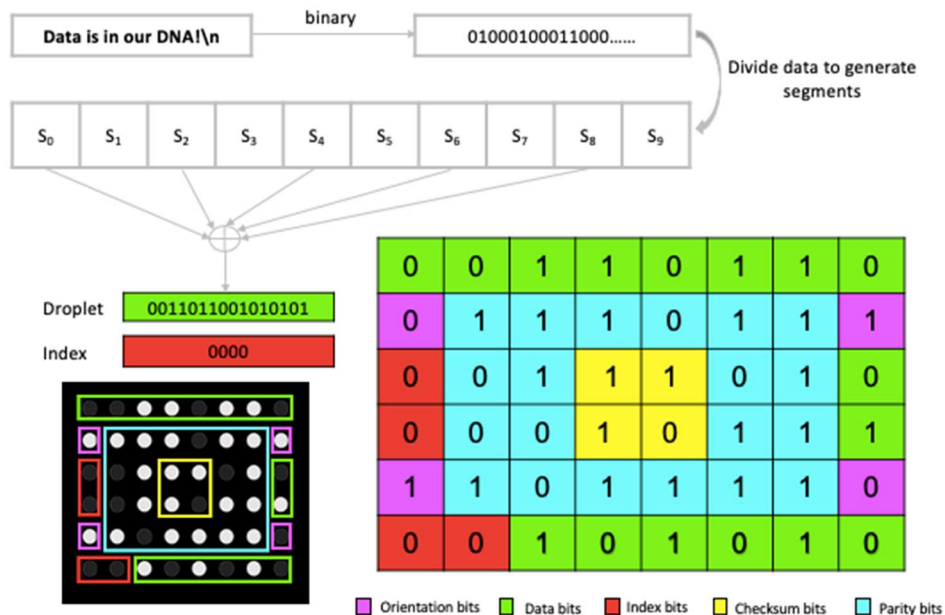


Figure 1. Example of Fountain Code implementation of dNAM digital encoding. The figure illustrates some of the main steps involved in encoding a digital message into dNAM. First a character string is divided into non-overlapping segments. These segments are combined in various patterns via an XOR operation to generate data droplets. Each droplet is assigned an index, error-correcting (checksum and parity) and orientation information and positioned within a grid to form the design used to synthesize a dNAM origami.

1.2 Size efficiency of the encoding scheme. Simulations were run to determine the size efficiency of the encoding scheme, as well as its ability to recover from errors. As shown in **Fig. 2A**, the number of origami required to encode a message of length n increases roughly at a linear rate up to $n = 5000$ bytes of data. Larger message sizes require more bits to be devoted to indexing, decreasing the number of available data bits per origami, creating a practical limit of 64 kilobytes of data for the prototype described in this work. This limit can be increased, however, by increasing the number of bits per origami. To determine the ability of the decoding and error correction algorithm to recover information in the presence of increasing error rates, *in silico* origami that encoded randomly generated data, were subjected to increasing bit error rates. The decoding algorithm robustly recovers the entire message for all tested message sizes when the average number of errors per origami is less than 7.4 (**Fig. 2B**). At 7.4 errors per origami, the message recovery rate dropped to 97.5%, and as expected decreased rapidly with higher error rates (55%

recovery at 8.2 errors per origami, and 7.5% at 9 errors per origami). An important feature of our algorithm is that the origami recovery rate can be low (as low as 63%) and still recover the entire message 100% of the time.

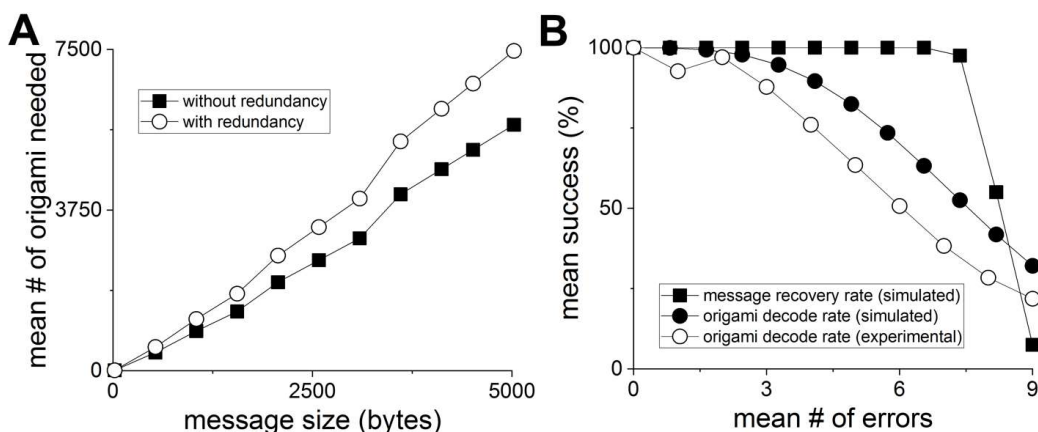


Figure 2. dNAM origami and message recovery rates in the presence of increasing errors. Simulations were performed to determine the theoretical success rates for correctly decoding individual dNAM origami and recovering encoded messages. The mean number of dNAM origami needed to successfully recover messages of increasing length with (open circles) or without (filled squares) redundant bits is plotted in (A). In (B) the mean success for recovering both individual origami (circles) and the entire message (squares) are plotted against the mean number of errors per origami (randomly generated for simulated data). Simulation recovery rates (filled symbols) are averages of all message sizes tested (160 to 12,800 bits). For experimental data (open circles) the mean success was estimated by comparing the decode algorithm’s results with that of the template-matching algorithm. Two types of dNAM origami were simulated, with (open circles) and without (black squares) redundancy.

1.3 Number of origami needed to decode the ‘Data is in our DNA!\n’ message. As a further test of the efficiency of the encoding/decoding algorithm, we used a random sampling approach to determine the number of origami needed to decode the ‘Data is in our DNA!\n’ message. We started with all the decoded binary output strings that were obtained from the single-field-of-view recordings and took random subsamples of 50-3000 binary strings. We passed each random subsample of strings through the decoding algorithm and determined the number of droplets that were recovered (Fig. 3). Based on the algorithmic settings used in the experiment, we found that only ~750 successfully decoded origami were needed to recover the message with near 100% probability. This number is largely driven by the presence of origami in our sample that were prone to high error rates and thus rarely decoded correctly.

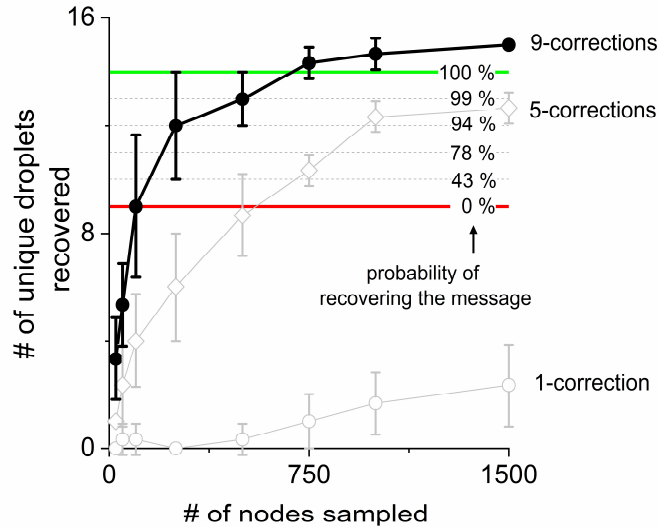


Figure 3. Number of dNAM origami required to recover the message. The mean number of unique dNAM origami matrices correctly decoded for randomly selected subsamples of decoded binary strings. The analysis was further broken out by the number of errors corrected for each origami, three examples are plotted (1, 5 and 9). Black filled circles depict the results for nine error corrections, which is the ‘maximum allowable number of errors’ parameter used in the decode algorithm for all other analysis reported here. The horizontal lines indicate the probability of recovering the message with different numbers of unique droplets. With fourteen or more droplets, the message should always be recovered (thick green line, and above indicates 100% chance of recovery) and with nine or fewer droplets the message will never be recovered (thick red line and below indicates 0% chance of recovery). Mean values for three experiments are shown. Error bars indicate \pm SD.

Task 2 – Create a high-throughput, integrated analytical engine to design select data strands using quantitative metrics based on an in-house, algorithm.

This task was completed during the prior review. The sequence selection software called SeqEvo has been made publically available and Dr. Mike Tobiason, who created the software during his PhD, has returned to Boise State as a postdoctoral fellow uses the software to design and select DNA sequences for Tasks 3, 4, and 5. We have recently purchased high-performing computational resources in reduce the time and to increase the scale of the sequences that we can design/select.

Task 3 – Create a synthetic biological factory for manufacturing DNA scaffolds using a rapid design, build, and test cycle of genomes.

Microscopy results and modeling indicated that the outside edges of the origami tend to be poorly resolved and less stable than the inside of the origami. We hypothesized that enlarging the origami would allow the same amount of data to be stored while avoiding the outermost edge of the structure. However, this would require a larger, custom-designed ssDNA scaffold. Toward this goal, 10 novel scaffold sequences were designed that were ~11 kb in length, more than 50% larger than the common M13 scaffold. Having multiple designs will allow us to test our hypothesis that larger origami can improve our structures in general, ensuring that results are not based on individual sequences. A model of the large origami was built, and this was fed into a monte-carlo based computational algorithm that designs scaffold sequences compatible with the restraints of this design (**Fig. 4**). These sequences were computationally inserted into plasmids to design the

DNA sequences that must be synthesized. A database of ssDNA scaffolds and plasmids is under development. Gene synthesis and ssDNA production protocols are also in development. This effort was led by PhD candidate Sarah Kobernat, and several undergraduate VIP students were involved in the computational-based engineering effort and data management. Preliminary results were presented at the 26th International Conference on DNA Computing and Molecular Programming (DNA 26, September 14–17, 2020).

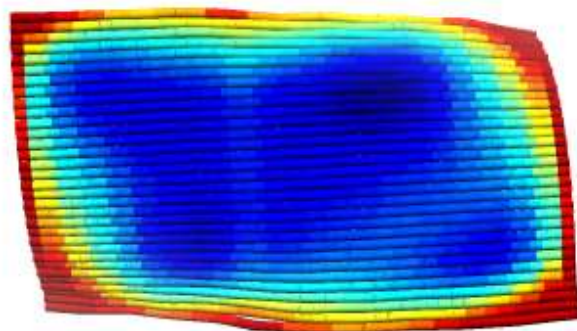


Figure 4. Origami edges are suboptimal for nanometer scale localizations. CanDo representation of an origami rectangle built from a 10.2kb custom DNA origami scaffold. Colors indicate the computed "flexibility" of the structure, where red is more flexible and blue is less flexible.

Task 4 – Design and fabricate NAM storage platforms using the DNA scaffolds, and validate the functionality of genome scaffolds using atomic force microscopy.

A workflow was developed for the design and fabrication of DNA origami storage nodes while developing the dNAM prototype. As described in *Task 1.1*, custom software was developed to encode digital data into patterns of data strands spread across multiple nodes. Additional software and excel spreadsheets were created to automate the selection of oligonucleotides, both for ordering and pipetting using an epMotion liquid handling system. Once origami were synthesized, quality control was carried out using DNA-PAINT SRM and atomic force microscopy. DNA-PAINT imaging indicated that, although the edges of origami were more sensitive to data strand insertion failures (**Fig. 5**), all of the data domains, in each of the origami designs, were detectable in each of three separate experiments (see *Task 5*).

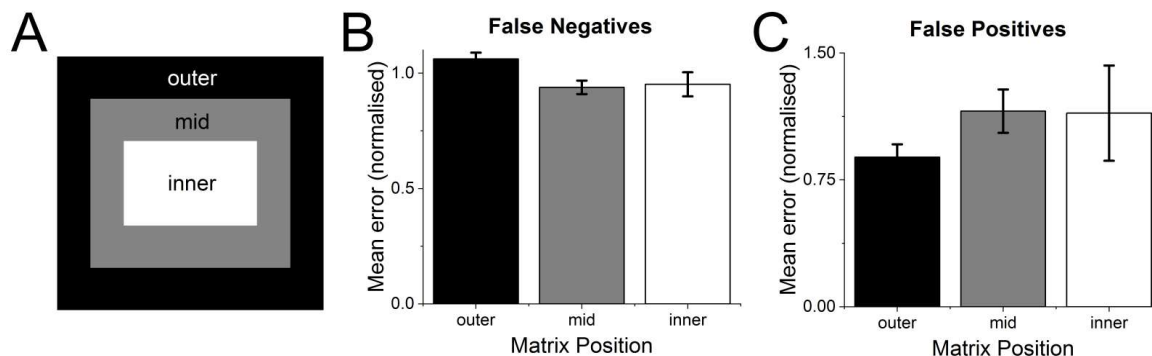


Figure 5. Outer edges and inner regions of dNAM origami are differentially error prone. The array positions of DNA origami (only considering structures with 15 or less errors, as identified by template matching) were classified as either ‘outer’, ‘mid’ or

‘inner’ depending on their position in the array (A). The mean error for each classification was calculated and normalized by dividing by the overall mean error for that zone. Plots of the mean normalized false negative (B) and false positive values (C) for each zone are shown. Mean values for three experiments are shown. Error bars indicate \pm SD.

Each individual origami synthesis was visualized and validated by AFM. The AFM images further confirmed that the general shapes of all 15 origami designs were as expected with properly positioned data domains (Fig. 6). The results indicate that the extended staple strands do not substantially inhibit the synthesis of the 15 unique origami designs.

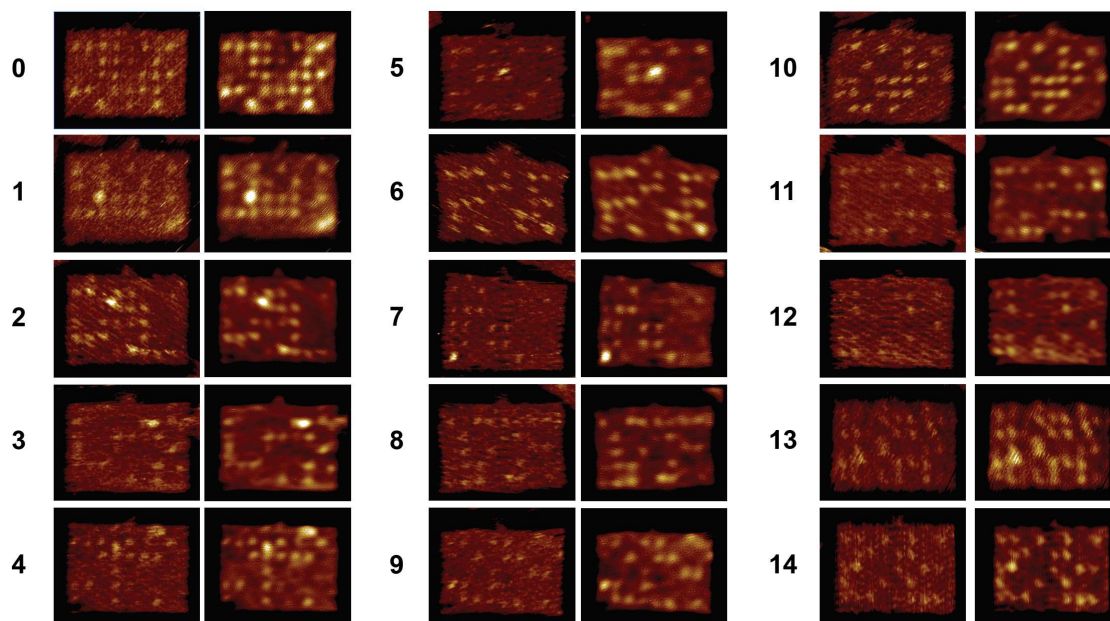


Figure 6. AFM images of dNAM origami. Representative AFM images of all 15 dNAM “Data is in our DNA! \backslash n” origami, where most of dockings sites are visible. (An inverse FFT analysis with a band rejected filter has been applied to highlight the docking positions in right-hand panels). Every image is 90 x110 nm² and the color scale ranges over 250 pm.

Task 5 – Read arbitrary data files into NAM storage nodes using super-resolution microscopy.

5.1 Successful dNAM prototype. To test our dNAM concept, we encoded the message ‘Data is in our DNA! \backslash n’ into 15 distinct DNA-origami nanostructures (Fig. 7A). Each origami was designed with a unique 6 x 8 data matrix that was generated by our encoding algorithm with data domains positioned \sim 10 nm apart. For encoding purposes, the message was converted to binary code (ASCII) and then segmented into 15 overlapping data droplets that were each 16 bits. Inspired in part by digital encoding formats like QR-codes, the 48 addressable sites on each origami were used to encode one of the 16-bit data droplets, as well as information used to ensure the recovery of each data droplet. Specifically, each origami was designed to contain a 4-bit binary index (0000 – 1110), twenty bits for parity checks, four bits for checksums, and four bits allocated as orientation markers (Fig. 7B). To fully recover the encoded message, we synthesized each origami separately, deposited an approximately equal mixture of all 15 designs (\sim 20 femtomoles of total origami) onto a glass coverslip, and recorded 40,000 frames from a single field of view using DNA-PAINT (\sim 4500 origami identified in 2,982 μ m²). Super-resolution images of the hybridized imager strands

were reconstructed from signal blinks identified in the recording to map the positions of the data domains on each origami (**Fig. 7C**). Using a custom localization processing algorithm developed by Dr Will Clay, the signals were translated to a 6 x 8 grid and converted back to a 48-bit binary string—which was passed to the decoding algorithm for error correction, droplet recovery, and message reconstruction. The process enabled successful recovery of the dNAM encoded message from a single super-resolution recording.

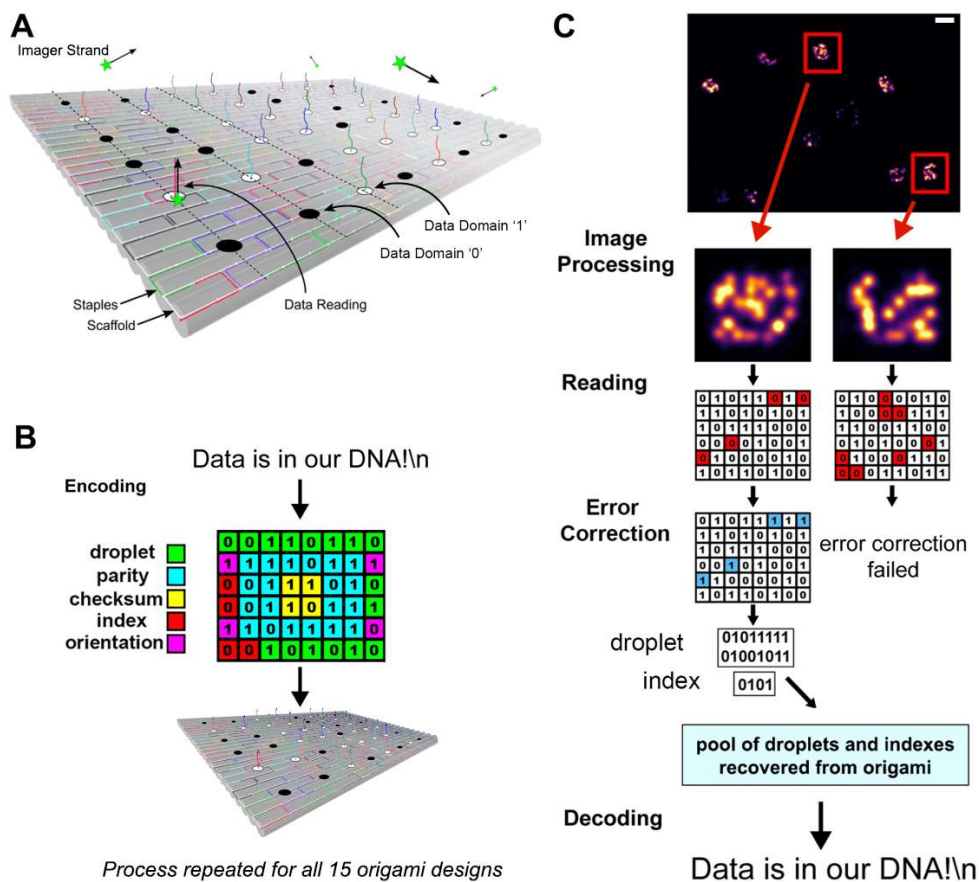


Figure 7. Binary dNAM overview. (A) Illustration of a binary dNAM origami, a DNA nanostructure with specific sequences used to localize data strands (a.k.a. information-bearing particles) to programmable sites within the DNA origami. Site-specific localization is enabled by extending/not-extending the structural staple strands of the origami to create physical representations of 1s/0s. The presence, absence, and identity of a data strand’s docking sequence defines the state of each data strand, and is assessed by monitoring the binding of data imager strands via DNA-PAINT. (B) To enable reading of our test message, ‘Data is in our DNA!n’, 15 dNAM origami were synthesized based on designs generated by the encoding algorithm (see Encoding in main text). For clarity, only one of the 15 designs is shown here. The colors of the matrix sites depicted in the design correspond with the roles of the site’s bit values as follows: droplet (green), parity (blue), checksum (yellow), index (red), and orientation (magenta). (C) To ‘read’ the message, 4 μ L of the DNA-origami mixture, containing 0.33 nM of each origami, was imaged using DNA-PAINT (top panel). The origami in the rendered image were identified and converted to an array of 1’s and 0’s corresponding to the pattern of localizations seen at each matrix location. The decoding algorithm performed error correction where possible, and successfully retrieved the entire message when sufficient data droplets and indexes were recovered. Scale bar, 100 nm.

As a quality control step, we evaluated all of the origami structures in order to confirm that the 15 different designs were successfully synthesized, with data domains in the intended addresses.

Automated image processing algorithms were developed to identify, orient and average multiple images of each origami from the DNA-PAINT recording of the mixture (**Fig. 8**).

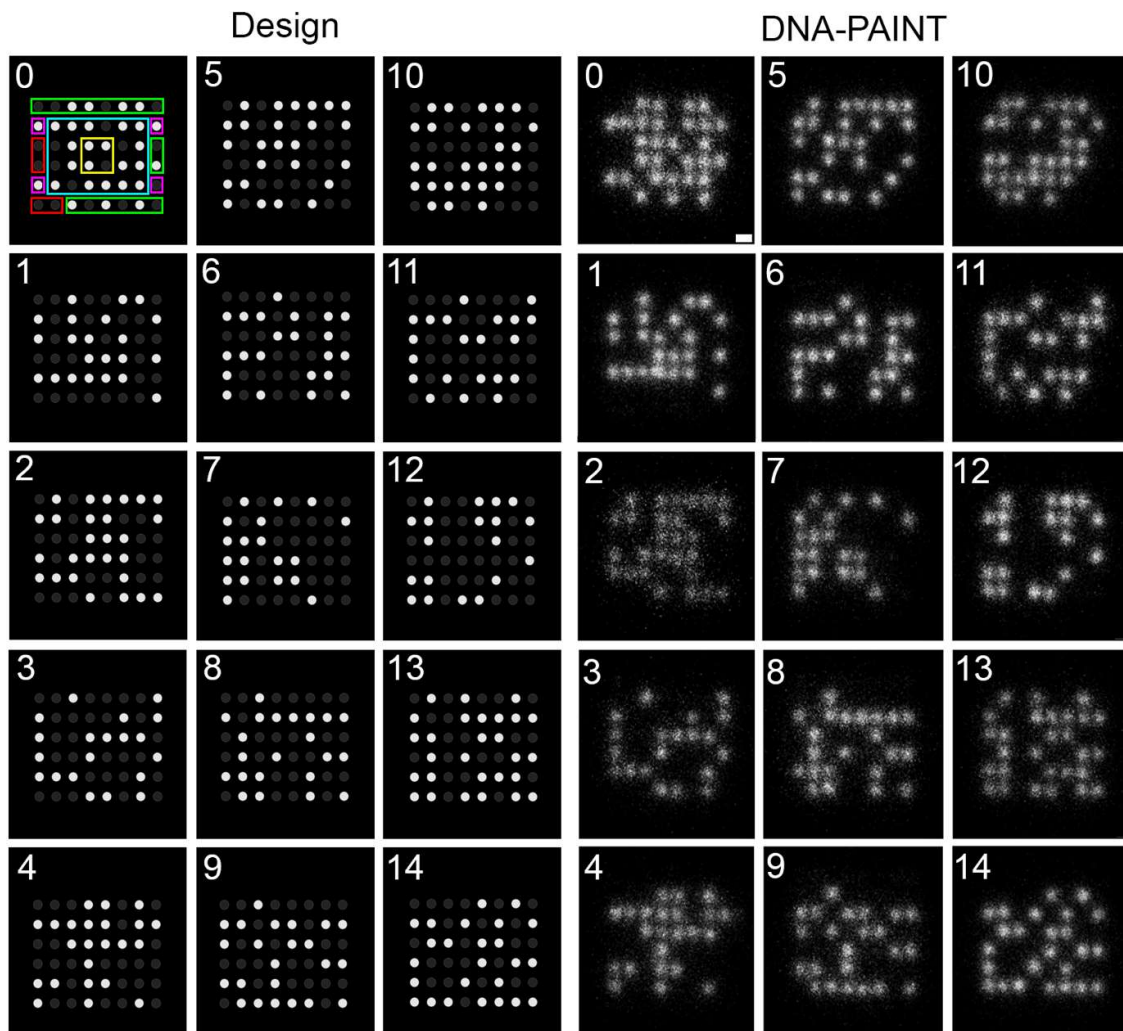


Figure 8. DNA-PAINT imaging of dNAM indicates all sites are recovered in a single read. dNAM origami from a DNA-PAINT recording were identified and classified by aligning and template matching them with the 15 design matrices (**Design**) in which all potential docking sites are shown, filled circles indicate sites encoded ‘0’ (dark grey) or ‘1’ (white). Colored boxes indicate the regions of the matrices used for the droplet (green), parity (blue), checksum (yellow), index (red), and orientation (magenta). For clarity, only the first design image includes the colored matrix sites. ‘Averaged’ images of 4560 randomly selected origami, grouped by index, are depicted right (**DNA-PAINT**). Scale bar, 10 nm.

Upon success, the mean number of each origami detected during a recording (**Fig. 9A**), the mean number of total errors including false positives and false negatives (**Fig. 9B**), the percentage of each origami successfully decoded (**Fig. 9C**), the percentage of each origami decoded versus the mean number of errors for each origami (**Fig. 9D**), and a comparison of the mean number of errors found in origami identified by template matching and the decode algorithm were calculated (**Fig. 9E**).

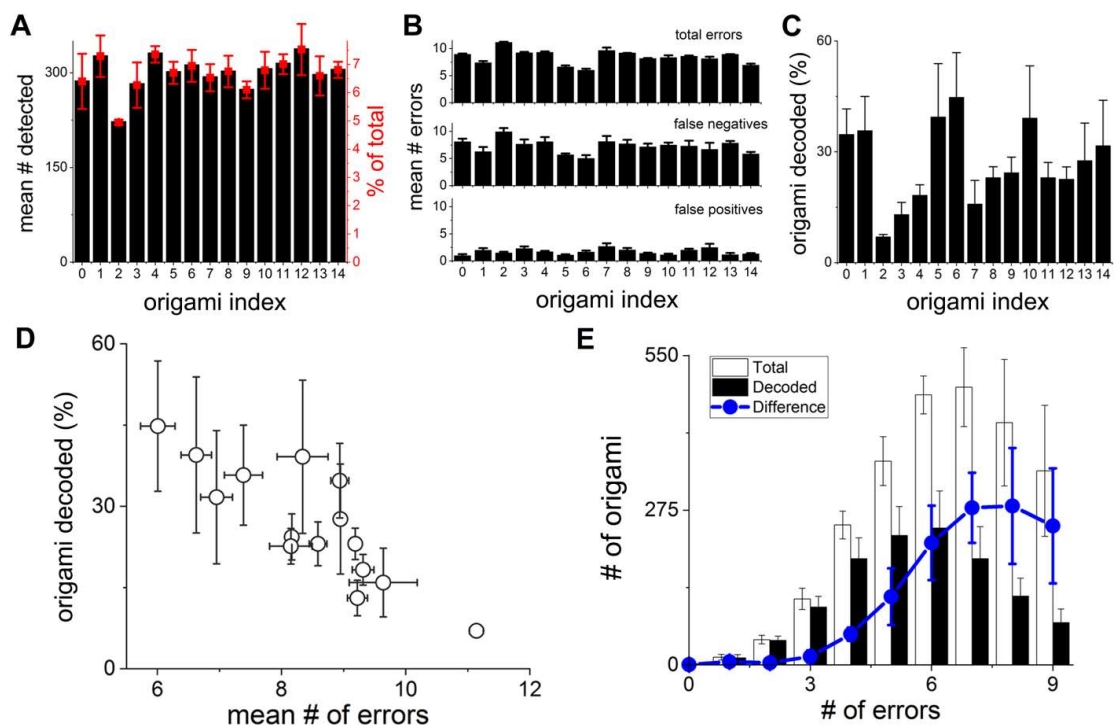


Figure 9. All 15 dNAM data strings were recovered from a single read. (A) plots the numbers of each index observed in a single recording, based on template matching. The mean counts are shown as black bars, percentage of total dNAM origami are shown in red. In (B) the mean number of total errors (top) for each structure is shown, based on template matching. The same errors are also shown after being grouped into false negatives (middle) and false positives (bottom). (C) depicts the percent of origami passed to the decode algorithm that had both their indexes and data strings correctly identified. In (D) the percentage of each origami decoded is plotted against the mean number of errors for each structure. (E) Histograms of the total mean numbers of errors found in origami identified by template matching (open bars) and the decode algorithm (black bars) are shown. The difference between the two is plotted in blue. Mean values for three experiments are depicted in all graphs, error bars indicate \pm SD.

5.2 DNA-PAINT resolution. To evaluate the resolution of the DNA-PAINT experiments used in the dNAM proof-of-principle, FWHM values were derived by taking transect measurements centered on binding sites in rendered images (with 1-pixel blur applied) of either individual or ‘averaged’ dNAM origami (Fig. 10). In both cases at least ten binding sites were examined for each structure using with horizontally or vertically aligned positioned transects (Fig. 10 A,B). FWHM values of $6.6 \text{ nm} \pm 1.6 \text{ SD}$ (single origami images, $n = 124$) and $7.2 \text{ nm} \pm 0.3 \text{ SD}$ (averaged origami images, $n = 47$) were calculated from Gaussian fits to plots of the transect data (Fig. 10 C,D). This is important because the experimental resolution limits the information that can be read, not stored, in nucleic acid memory.

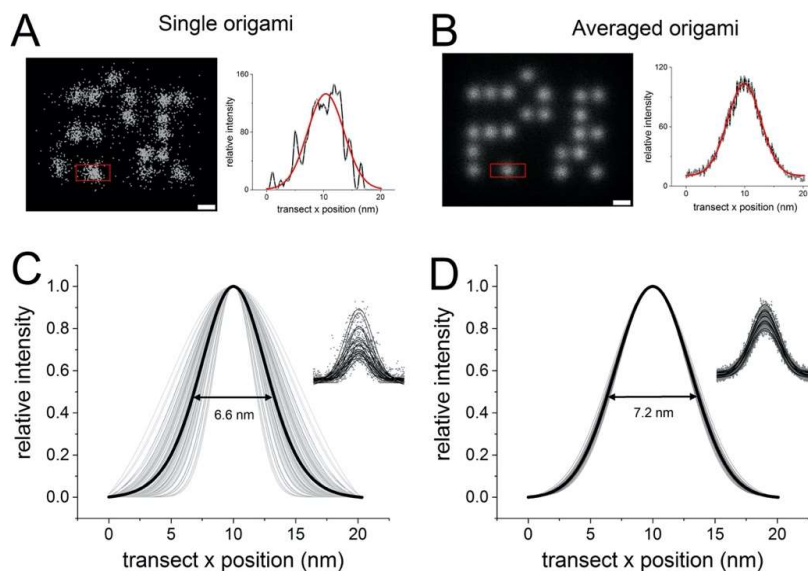


Figure 10. DNA-PAINT imaging demonstrates low nm resolution. FWHM values were derived by taking transect measurements centered on binding sites in rendered images of either individual (A) or ‘averaged’ dNAM origami (B). Transects were placed horizontally (as shown in red) on vertically for measurements. A plot from a single binding site is shown with a Gaussian fit to the data plotted in red. Gaussian fits for binding sites from each experiment are plotted in grey for both single (C) and ‘averaged’ (D) structures (after centering and normalization). The mean of the grey lines is shown in black. The inset plots are the representative results from a single experimental recording. The mean FWHM value for individual fits to each experiment was calculated and reported in the main text. Origami-6 was used in all cases, as it was the most consistently recovered structure. Scale bars, 10 nm.

5.3 – Sub-nanometer imaging resolution for SRM. Currently, we are capable of achieving a maximum resolution of ~ 5 nm using DNA-PAINT SRM. One route for further improvements in resolution is sample stabilization. To this end, Dr Will Clay has developed a method for real-time, active drift correction of a microscopy stage using tracking of multiple fiducial markers during acquisition. The system uses the same illumination optics, imaging optics, and sensor to detect molecular localizations and track the position of the markers so it is able to account for all potential sources of signal drift without any additional alignment or stabilization. The system uses a piezo nano-positioning stage to correct the sample drift between frames. The system is capable of stabilizing the position of the fiducial markers to less than 0.5 nm root mean square error when they are imaged alone and to 0.9 nm when they are imaged in the presence of many single molecule emitters. A visualization of this method is shown in **Figure 11**. The method identifies as many as 80 fixed markers on the sample and tracks their position using emitter localization in every frame in real time. The large number of trackers allows for averaging many positions, allowing for sub-nanometer measurement of drift even when the individual markers have poor signal to noise. **Figure 12** shows the results of a DNA-PAINT measurement taken on the same field of view of the same sample without (left) and with (right) our active drift correction system operating. Neither image has been corrected with post-processing drift correction. Without correction, the image is blurred significantly with long streaks, as is typical for uncorrected PAINT imaging. Post-processing analysis of this un-corrected image revealed over 1 micron of drift, indicating that the

images would have drifted over 10 camera pixels during the course of the experiment. With active drift correction, the image was stable enough to resolve the emitting sites on the individual origami, spaced 20 nm apart, without any post processing. **Figure 13** shows the results of post-processing analysis on the actively corrected image to determine the amount of residual drift in each frame. The left panel shows the residual error in each of the 1000 frames collected, demonstrating little structure to the remaining drift and showing that the stage was stabilized to ± 4 nm in over 99% of the frames and ± 7 nm in every frame. Note that each frame is 300 ms long so 1000 frames spanned a 5-minute experiment. The right panel shows the same data collected as a histogram of the residual error. The RMS width of this distribution is 0.9 nm. The average number of localizations identified in each frame was roughly 2000, enough to form a DNA-PAINT image. Similar analysis of data taken on the same sample before the fluorophore-labeled DNA was introduced showed 0.5 nm RMS residual error, indicating that the systems performance is impacted by the presence of the DNA-PAINT blinks but not severely.

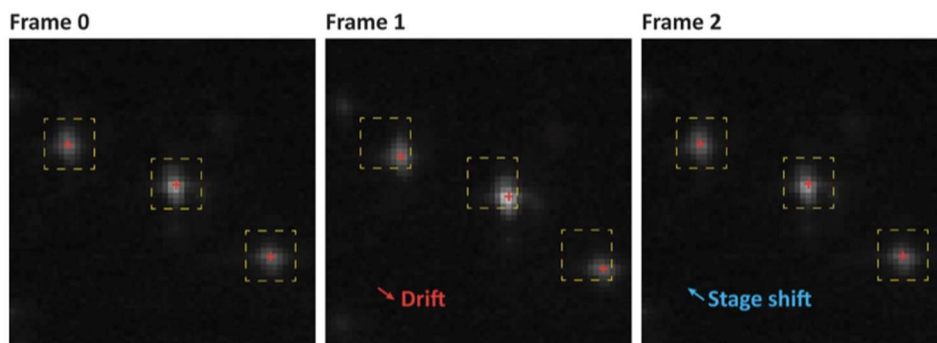


Figure 11. Schematic representation of the drift correction system. Yellow rectangles show regions of interest (ROIs) containing the signal from a fiducial tracking marker attached to the microscope slide. Red crosses show the fitted center position of the signal. In Frame 0, the markers are located in the center of the ROIs. In Frame 1, the positions have moved due to drift. In Frame 2, the stage has shifted to restore the markers to their original positions. The system applies these corrections after every frame with a minimum step size of 0.4 nm.

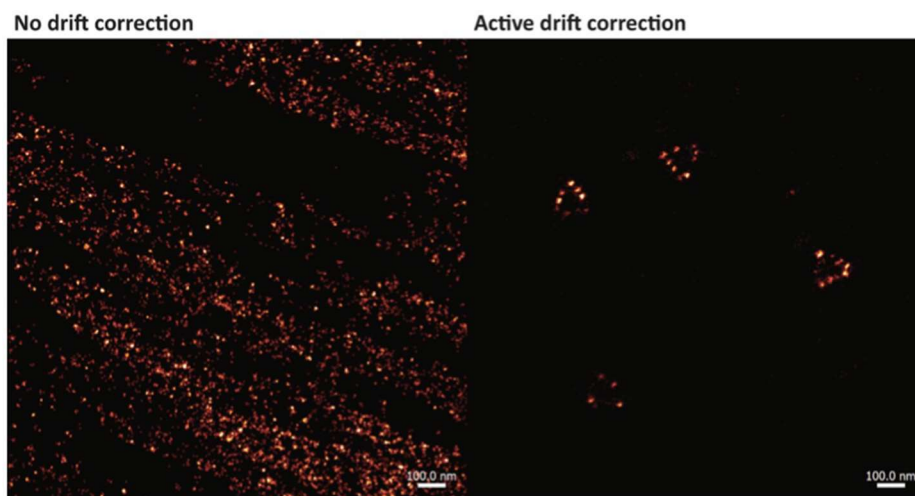


Figure 12. Super Resolution Microscopy Drift. Localization data collected from a super-resolution image of triangle shaped DNA origami, without any drift correction (left panel) and with active drift correction but no additional post processing (right panel). Resolved docking sites on the origami are spaced approximately 20 nm apart.

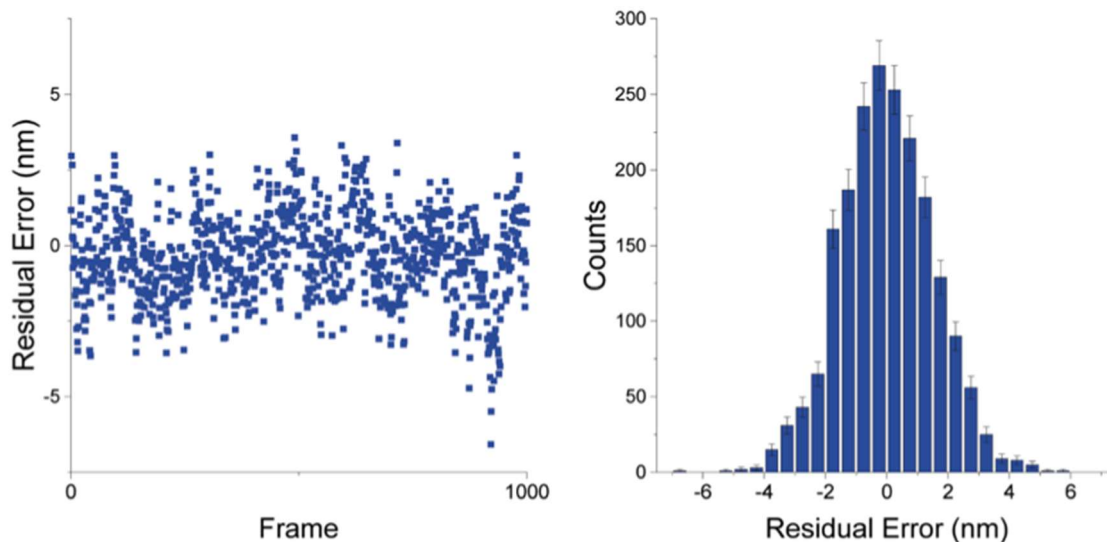


Figure 13 Error Analysis. Residual error in localization positions caused by uncorrected drift, based on post processing analysis of the image. Left panel shows the residual error in each of the 1000 frames of the experiment. Right panel shows same data in a histogram showing the occurrence of different amounts of error. The RMS width of the distribution is 0.9 nm.

IV. Demonstration of economic development and impact

Demonstration of Economic Development and Impact	<u>Year 1 Reporting Period</u> 07/01/2018–06/30/2019	<u>Current Reporting Period</u> 07/01/2019–12/31/2020
External Funding	\$ 1,549,995	0
Number of External Grants	3	1 submitted, not recommended for funded
Private Sector Engagement	14 companies	2 companies, 1 VC group
University Engagement	11 universities	~20 universities
Federal Agency Engagement	5 agencies	4 (NSF, SRC, NRL, NIST, IARPA)
Industry Involvement	2 companies	2 companies (Micron, SRC)
Patents	0	1 Patent Application
Copyrights	0	0
Plant Variety Protection Certificates	0	0
Technology Licenses Signed	0	0
News Releases	3 articles	0
Start-up Businesses Started	0	1
Jobs Created outside of Boise State	0	6

External Funding

During this reporting period, we pursued the NSF SemiSynBio II proposal opportunity to grow dNAM from a 2D to a 3D technology. We were not awarded in part because of our active SemiSynBio I Award, as well as concerns that our next generation ideas were too ambitious. The benefit of applying is that our team is better positioned to resubmit the proposal next year and our team is clear minded on the preliminary results that we need to secure to achieve this goal. Target opportunities during the next award period include the NSF SemiSynBio III (if available), the NSF Partnerships for Innovation, and the NSF Planning Grants for Engineering Research Centers.

Engagement

With the backing of the NSF Office of Emerging Frontiers and Multidisciplinary Activities, Hughes hosted the 2019 Germination Meeting at Boise State University on August 15-16, 2019. The meeting focused on new approaches in cultivating risk-taking and impact-driven research culture. In response to this event, Hughes was encouraged by NSF to co-create an Institute for Transformative Scholarship at Boise State University to help researchers overcome the structural, cultural, personal, and financial barriers that prevent them from germinating and pursuing bolder ideas. As noted in the Year 1 Annual Report, the National Science Foundation (NSF) in collaboration with the Semiconductor Research Corporation (SRC) jointly awarded the research team \$1,500,000 to address the scientific challenges facing NAM technologies. The funding mechanism was called *Semiconductor Synthetic Biology for Information Processing and Storage Technologies*. As part of this funding, the SRC holds an annual conference to showcase “*the quality and breadth of the SRC research portfolio, the excellence of SRC students and faculty, and the magnitude of the collaborative research investment made by industry through SRC.*” Hughes and two PhD students (now graduated) on the NAM team, Chris Green and Mike Tobiason, attended the conference, which was held in Austin, Texas from September 8–10, 2019. “*The conference features student presentations and posters and gives SRC member companies multiple formal and informal occasions to network with SRC students. This is a great opportunity for students to meet with SRC member companies, including 7 of the top 10 semiconductor companies in the world. These networking occasions with SRC member companies give student opportunities to open the door to future full-time employment.*”

Hughes was also among a select group of scholars, industry stakeholders, and program managers to participate in a workshop on Nucleic Acid Nanotechnology. The workshop, held in Boston, Massachusetts in December 2019, was co-sponsored by the Materials Research Society and the prestigious Kavli Foundation. The goals of the workshop were to establish “priority research areas for next-generation applications of nucleic acid nanotechnology across diverse domains spanning computation, sensing, fabrication, therapeutics, and other areas.” Through this process, Hughes reinforced relationships with Harvard University (George Church, William Shih), MIT (Mark Bathe), NIST (James Liddle), NRL (Igor Medintz), John’s Hopkins (Rebecca Schulman), as well

as established new relationships with the editors of Nature and Nature Materials. Based on ideas shared, George Church opened his lab to members of the NAM Institute at Boise State.

Business Development

Steven Burden, who successfully completed his PhD in Biomolecular Sciences, graduated

Classification	Number
Tenured or Tenure Track Faculty	4 (<i>2 full professors, 2 associate professors</i>)
Research Faculty	1 (<i>started a tenure-track faculty position</i>)
Project Manager	1 (<i>also focused on business development</i>)
Senior Lab Research Associate	1 (<i>manages the laboratory & supports team</i>)
Postdoctoral Fellows	3 (<i>performing at a research faculty level</i>)
Graduate Students	6 (<i>3 of the 6 graduated in December 2019</i>)
Undergraduate Students	10 (<i>5 female and 5 male</i>)

December 2019 as a member of the NAM Institute. Burden’s dissertation topic was on the development of nucleic acid biosensors with allosteric fluorescence signals. For the NAM team, Burden played a lead role in our Vertically Integrated Project (VIP), where he trained undergraduate students to produce, purify, and ensure the quality control of single-stranded DNA scaffolds. Prior to graduating from Boise State, Burden co-founded a biotechnology startup (FACible BioDiagnostics — <https://www.facible.com/>). Based in Boise, Idaho, FACible BioDiagnostics is focused on developing rapid, low-cost, diagnostics. Burden began full time employment as the company’s CEO on January 1, 2020. In addition, one of the co-founders, Clementine Gibard Bohachek, was a postdoctoral research scientist at Boise State University and was part of the NAM team during the spring of 2019, where she developed VIP training materials and trained VIP and NAM graduate students on practical laboratory approaches to synthetic biology. In all, FACible BioDiagnostics employs 6 people — three full time and three part time. The financial, scientific and professional support that Burden received during his PhD was critical for his ability to secure venture capital needed to start his company. The success of Burden highlights the entrepreneurial environment that is being cultivated by the NAM Institute and team. The team is actively reflecting on attempting to spin-off a second company related to the project.

V. Numbers of student, staff, and faculty participation

From a professional development perspective, the goal of the NAM Institute is to ensure the success of the people that make up the team, from students and postdoctoral research scientists to the faculty and staff that enable open innovation, ideation, and collaboration. And with any academic environment, matriculation to graduation is expected, supported, and applauded. Thus,

we are happy to report that during this reporting period three PhD students on the NAM project successfully defended their PhD dissertation and graduated:

- **Steven Burden**, PhD in Biomolecular Sciences, Dissertation — *The Development of Nucleic Acid Biosensors with Allosteric Fluorescence Signals*
- **Chris Green**, PhD in Materials Science and Engineering, Dissertation — *Nanoscale Optical and Correlative Microscopies for Quantitative Characterization of DNA Nanostructures.*
- **Mike Tobiason**, PhD in Materials Science and Engineering, Dissertation — *Engineering Kinetically Uniform DNA Devices*
- **Golam Md Mortuza**, successfully completed their PhD proposal in Computer Science.

In addition, Reza Zadegan has started a tenure track faculty position at North Carolina A&T this past August. His professional development included but was not restricted to: grant writing support by Watson and Hughes; germination of research directions and intellectual risk management by Hughes; helping create his faculty package by Hughes, Andersen, and Hayden; mock interviews by Hughes; national and international networking opportunities by Hughes; technical training by Andersen, Hayden, Kuang, and Hughes; and professional training from Hughes and Watson. We also would like to acknowledge that one of the project principle investigators, Elton Graungnard, has transitioned from the team to focus his efforts on developing atomically-thin semiconducting materials for high performance, energy-efficient electronic devices.

New Hires

During this reporting period, two postdoctoral research scientists were hired: Luca Piantanida, who started on August 5, 2019 and Mike Tobiason, who started on November 16, 2020. Piantanida has a PhD in Nanotechnology from University of Trieste, where his dissertation was on developing DNA origami nanoactuators functionalized with gold nanoparticles for plasmon resonance tuning. Piantanida recently concluded a postdoctoral position at Durham University, UK under the supervision of Prof. Kislun Voitchovsky, where he developed a novel atomic force microscopy approach for imaging biological interfaces in fluid. The no-cost extension provided by IGEM-HERC for Year 2 funding enabled the team to hire Mike Tobiason (he is being supported through Year 3 IGEM-HERC funding). Tobiason was a previous PhD student in the NAM Institute having earned his PhD on his work titled, “Engineering Kinetically Uniform DNA Devices.” Tobiason’s DNA biotechnology expertise along with his deep knowledge of NAM, he has been able to make an immediate impact to the project. In addition to the postdoctoral research scientists, we also hired a new Ph.D. student, Sarah Kobernat, in support of designing, producing, and optimizing large DNA origami scaffolds (Task 3). She is also supporting the Vertically Integrated Project through mentoring undergraduate students on scaffold design.

Vertically Integrated Project

The Vertically Integrated Project (VIP) model integrates teaching and learning into one framework in support of work-force development of students that can work at the interface of semiconductor

manufacturing and synthetic biology. These students are engaging in research activities aimed toward the production, purification, and quality control of new single-stranded DNA origami scaffolds. The students range from sophomore to seniors and span four different majors: biology, chemistry, health sciences, and psychology. Specifically, the VIP students synthesized and purified several large DNA scaffolds. During the Fall 2019, the VIP students used *E. coli* cultures to express single stranded DNA ~8,000 and 10,000 bp in length. Currently, the bacteriophage M13mp18 is used to make the DNA scaffolds, but it limited to 7249 nucleotides. In addition to being longer than M13mp18, each of these scaffolds has a different sequence, potentially enabling orthogonal origami synthesis. In the Spring 2020, due to COVID restrictions, the VIP students transitioned from synthesis and characterization work in the laboratory to scaffold design and modelling work as described in *Task 3*.

VI. Description of Future Plans

Team Management – Integration and graduation

- Manage the financial risk of the anticipated higher education budget cuts in Idaho, in response to COVID-19, that have the potential to impact the NAM Institute.
- Target the next round of grant opportunities and start working towards their submission. Leverage the grant writing process as an opportunity for professional development of the postdoctoral fellows. These include the NSF Major Research Instrumentation Program, the NSF Partnerships for Innovation Program, potential SRC research development avenues, and the NSF SemiSynBio-III.
- Help the postdoctoral fellows identify the intellectual space they want to lead in the future; periodically meeting with them to establish their professional development plans.
- Seek collaborations with key internationally recognized research groups; with an eye for cross-training our laboratories.

Task 1 – Create improved algorithms for coding information into data strands.

- Use a convoluted encoder-decoder network to attempt to obtain a super-resolved image directly from the raw blinking events. While this technique has already been demonstrated on STORM (<https://github.com/EliasNehme/Deep-STORM>), when applied to our application space, it would be faster and could handle denser data better.
- Sharpening or deconvolving super-resolved images via a CNN trained on origami. There are multiple examples of successful applications of this technique to biological imaging

that may be more applicable to our system because of its rectilinear design (<https://csbdeep.bioimagecomputing.com/tools/care/>).

- Using object identification to automatically identify/classify origami in a super-resolved image to obtain binary strings to pass to our decoding algorithm. Our team has previously trained a YOLO-based CNN to identify buildings in aerial images (<https://arxiv.org/abs/2004.10934>), and it could be adapted for origami.

Task 2 – Create a high-throughput, integrated analytical engine to design select data strands using quantitative metrics based on an in-house, algorithm.

- The technical aspects of this task are complete. Next steps are to run SeqEvo on a high-performing cluster to decrease the time to run a job and increase the scale/complexity of the jobs that can be run.

Task 3 – Create a synthetic biological factory for manufacturing DNA scaffolds using a rapid design, build, and test cycle of genomes.

- With the successful development of software to optimize sequences, we will next set out to design and synthesize large scaffolds with sequences optimized for our specific origami designs. Several designs will be synthesized and compared. The super resolution microscopy advancements will aid in this comparison. This will require deeper integration from the research team.
- Develop quality control metrics for scaffolds. Each scaffold synthesis will need external quality control metrics to ensure batch to batch consistency in order to enable comparison.
- Determine the applicability of mass-spectrometry of DNA staple strands for defect analysis. We hypothesize that mass-spectrometry may provide information on several types of DNA damage that could lead to poor DNA-PAINT performance, such as depurination and deamination, that are not resolvable by other methods.

Task 4 – Design and fabricate NAM storage nodes using the DNA scaffolds.

- In addition to read and write using the dNAM platform, investigate editing.
- Explore the application of short Locked Nucleic Acid (LNA) and other DNA analogues in dNAM to increase the resolution of the super-resolution microscope during DNA-PAINT, as well as explore if $3D$ NAM (which is a derivative to seqNAM) is a viable system to increase information density in NAM.

Task 5 – Read arbitrary files into NAM storage nodes using super-resolution microscopy.

- Test methods to improve resolution on existing microscope, including reducing drift, improving drift correction, and increasing the signal-to-noise ratio.
- Use knowledge gained from optimizing existing microscope to design and test components for custom built super-resolution microscopy system while working toward a full prototype.
- Use simulation to better understand optimal imaging and sample design parameters to maximize data reading rate.

VII. Summary of Budget Expenditures

IGEM-HERC graciously allowed a no-cost extension of Year 2 funding through December 31, 2021. The below table summarizes expenditures associated with the project from July 1, 2019 to December 31, 2020. It also includes a budget adjustment made in response to the impact of COVID-19 on research operations. With the temporary closing of the research laboratories and the moratorium on travel, we projected \$37,000 would be remaining at the end of our Year 2 no-cost extension (Dec. 31, 2020). We requested that \$36,780.50 be re-budgeted into *Other Expenses* so the team can purchase DNA supplies in support of the research effort, which was approved by IGEM-HERC and \$219.50 be placed in *Student Fees* to cover an overage. *Salary* and *Fringe* supported our graduate students, postdoctoral research scientists, an assistant research professor, a laboratory manager, and a project manager. *Other Expenses* were used to purchase modified and unmodified DNA oligos, supplies to process modified and unmodified DNA oligos into dNAM, super-resolution microscopy supplies, atomic force microscopy supplies, and computers. The oligos are used to assemble NAM blocks and to perform super-resolution microscopy studies. The atomic force microscopy supplies complement the super-resolution studies by confirming the design and structural stability of the dNAM. The computers were purchased in support of our algorithm development and newest postdoctoral research scientist. Major *Capital* purchases include an upgrade to our epMotion system, which enables us automate the mixing of solutions to synthesize DNA origami in both an accurate and efficient manner. The system was malfunctioning and was approaching its end-of-life. The upgrade ensures vendor support throughout the life of this project. Capital was used to purchase a server to improve our image processing efficiency. As part of analyzing dNAM, we compile over 60,000 high resolution images (~40 GB) per experiment. Post-processing of each series of experiments, and the 60,000+ images, are computationally intensive. When performed on a desktop computer, processing requires hours to days of processing time per experiment. The server is enabling more efficient image analysis. We also purchased an enclosure for the super-resolution microscopy system to enable low noise imaging, which will enhance image resolution.

Category	Original Year 2 Budget	Year 2 After Budget Adjustments	Expenditures	Encumbrances	Remaining Budget
Salary	\$282,671.00	\$276,201.00	\$276,761.14	\$307.20	\$(867.34)
Fringe Benefits	\$96,375.00	\$83,138.88	\$83,570.50	\$63.20	\$(494.82)
Other Expenses	\$93,500.00	\$130,500.00	\$94,860.30	\$35,000	\$639.70
Travel	\$15,000.00	\$586.62	\$586.62	--	--
Capital	\$150,000.00	\$146,900.00	\$146,840.28	--	\$59.72
Student Costs	\$28,954.00	\$29,173.50	\$29,173.50	--	--
Total	\$666,500.00	\$666,500.00	\$631,792.34	35,370.40	\$(662.74)

VIII. Commercialization Revenue

Commercialization	Revenue
None.	\$0

IX. Additional metrics established specific to individual project

Metrics	Number
External Funding	\$ 1,549,995
Graduate Degrees Awarded	4 (3 PhD, 1 MS)
Dissertations Published	4 (3 PhD, 1 MS)
Invited Technical Presentations	15 (5 oral, 10 poster)
Software Tools Created	3
Peer-Reviewed Publications	1
Manuscripts in Preparation	4
Number of Graduate Students on the Project	2
Number of VIP Students Enrolled (grad and undergrad)	~10
Number of National and International Postdocs Hired	4
Number of Scientists that have become Tenure Track Faculty	1 (North Carolina A&T)
Number of PhD Students that have received Postdoc Fellowships	2 (NRC Fellowship)
Number of PhD Students that started a Company in Idaho	1 (6 employees)

Note: Listed above are specific, objective, measurable, and realistic performance metrics over the lifetime of the project. These metrics, many of which have been distributed throughout this report, are a reflection of project success and inform economic impact.