BOISE STATE UNIVERSITY

IGEM # 19-002: Nucleic Acid Memory

July 1, 2020 – January 1, 2021 Progress Report

Will Hughes

Tim Andersen

Eric Hayden

Wan Kuang

Will Clay

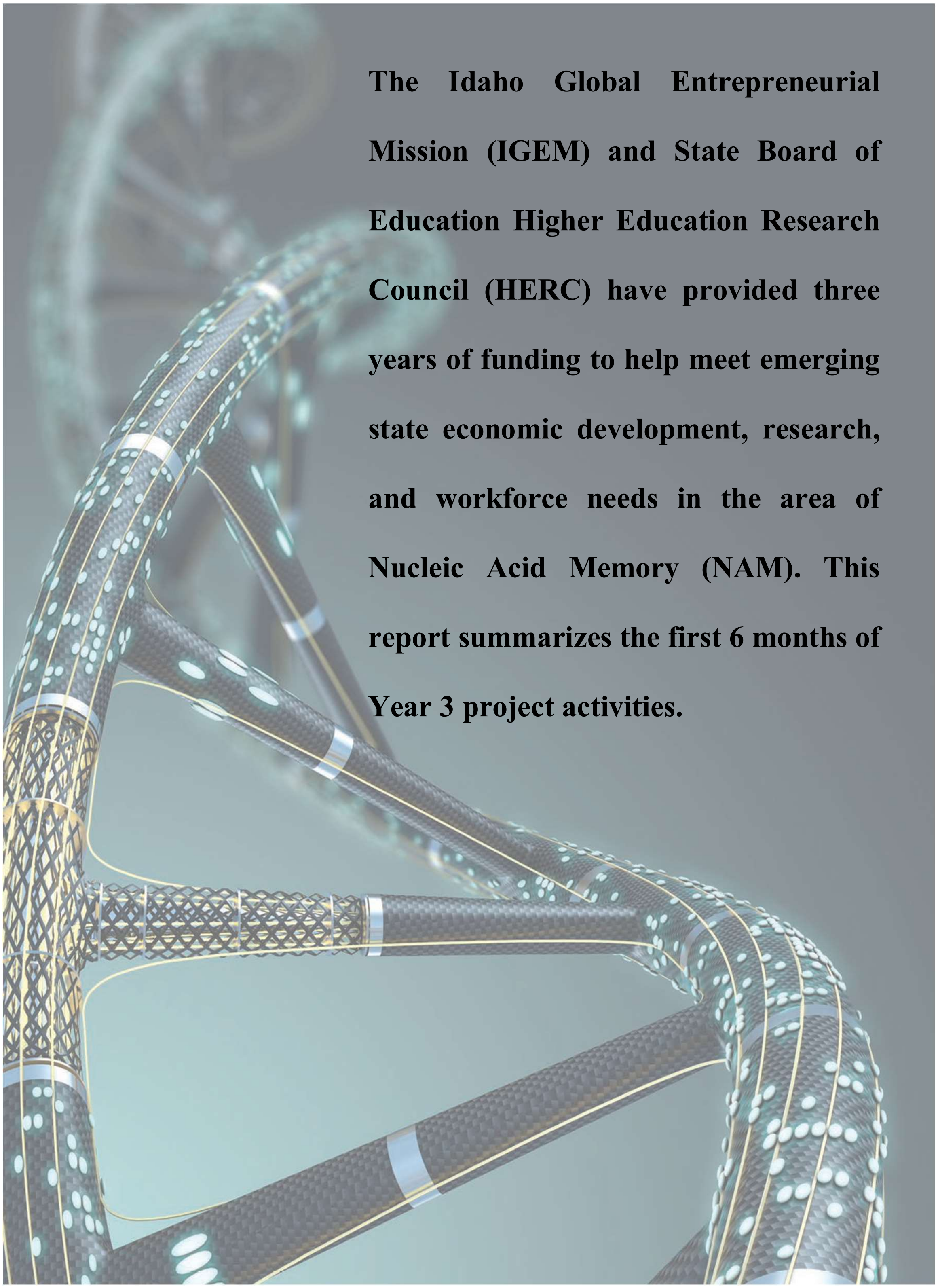
George Dickinson

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 the first 6 months of Year 3 project activities.

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, 2020 to January 1, 2021. 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 and year 2 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 prior Annual Reports — we conducted a series of experiments that validated dNAM as a platform for DNA-based information storage. The knowledge gained from this research resulted in the team generating new techniques and architectures for archival memory storage applications, which were further developed in response to the Semiconductor Synthetic Biology for Information Storage and Retrieval (SemiSynBio-II) request for proposals. While we were not selected for funding at the level of $1.5M, in part because we have an active grant through this funding mechanism, the ideas generated and the proposal development process were a galvanizing experience for the team that has led us to explore extending the dNAM platform supported by IGEM-HERC from two dimensions (dNAM and seqNAM) to three dimensions (3DNAM).

In brief, 3DNAM integrates time-correlated super-resolution microscopy and DNA self-assembly to digitally read non-volatile and randomly accessible information in all three dimensions (**Figs 1 and 2**). For 3DNAM, data is encoded by the presence or absence of hybridization events at *n* specific binding sites on a data strand. With 4 unique dyes (**Fig 1**), each site has the potential to encode 2 bits of digital information, producing 2*n* bits per strand, resulting in information densities above 10 Tbit/cm^2. More specifically, data strands extended from DNA origami nanostructures are modified with quenchers that enable lifetime measurements. The site-specific sequence of the data strands encodes the stored information and is read below the diffraction limit of light using

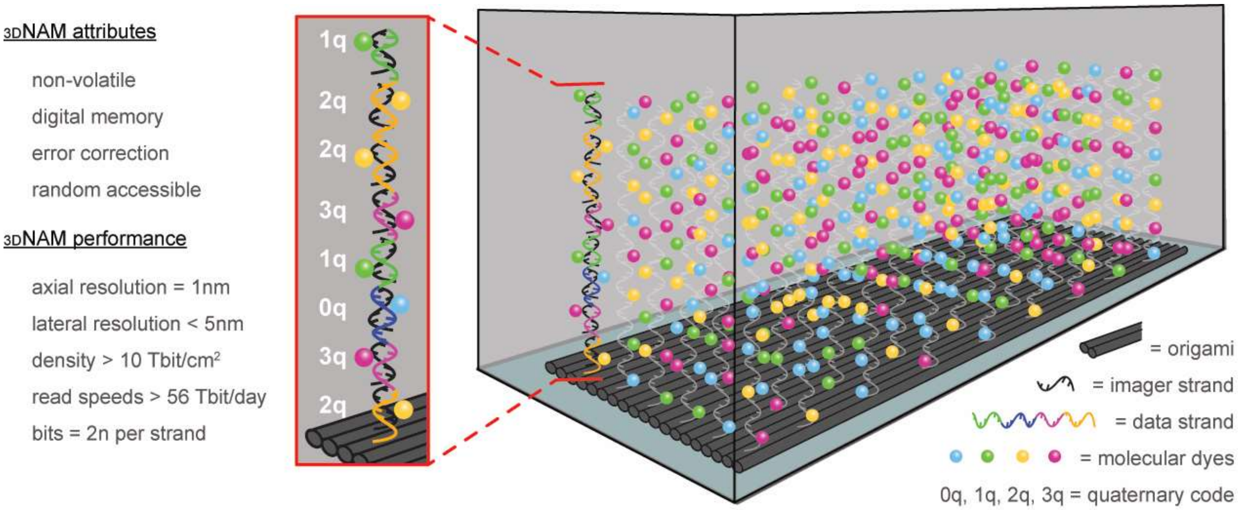


Fig 1. Schematic representing information density in three-dimensional nucleic acid memory (3DNAM). Colored spheres are dyes that reflect the digital information encoded into data strands (colored in inset) on DNA origami (dark grey) via transient binding of imager strands (black). 3DNAM has 2*n* bits per strand, where n is the number of binding sites on the data strand.

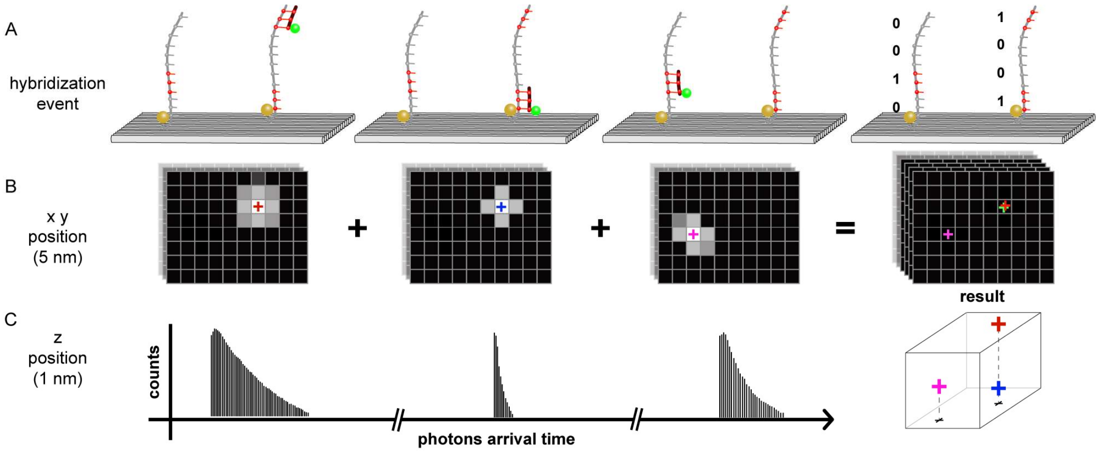


Fig 2. Illustration of 3DNAM read by TCSRM. The x-y position is localized from hybridization events between dye-labeled (green) imager strands and their sequence complements on a data strand (red). The z position is calculated from the lifetime fluorescence between a quencher (gold) and its dyes during FRET.

time-correlated super-resolution microscopy (TCSRM). TCSRM fully integrates DNA-PAINT super-resolution microscopy and fluorescence lifetime measurements to monitor interactions between dye-labeled imager strands and their sequence complement in three dimensions (**Fig 2A**). DNA-PAINT allows sub 5 nm lateral resolution of imager strand positions (**Fig 2B**), while fluorescence lifetime measurements provide the axial distance between imager strands and their respective quenchers with 1 nm precision (**Fig 2C**). TCSRM is superior to conventional SRM where the axial resolution of 40 nm limits the ability to read dense information in 3D. Similar to 3D flash memory, growing NAM in the 3rd dimension is an import memory innovation. Additional details on 3DNAM are provided in *Task 5.2* and *5.3*.

With an eye towards prototyping 3DNAM for future grant opportunities, the following report details the major work and outcomes supported by IGEM-HERC from July 1, 2020 to January 1, 2021, 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 Neural Network-based techniques to improve resolution and read errors in SRM images. In our prior work, we have successfully stored and recovered 20 bytes of information in our dNAM architecture. From this effort, we discovered that error correction bits consume more than half of the available 48 bits of encoding space on each origami. Out of those 48 bits of data, the error correction code used 28 bits of data, more than 58% of the total data space. No image averaging was performed, which makes the decoding process more error-prone as individual images are noisy. We are planning to train a neural network (NN) in support of optimizing image resolution and the decoding process. Since NNs are data hungry, a large amount of training data is required. Since it is not feasible to generate training data directly from SRM images, we are using the Picasso framework to simulate readout data. In doing so, we have written a training data generator to generate single dNAM images that are then randomly placed in a frame-by-frame sequence. After the dNAM images are generated from the Picasso module, noise is introduced. The noise is generated from the Poisson distribution.

1.2 Algorithms for encoding/decoding information onto seqNAM. The objectives of this sub-task are to develop (1) a more robust approach to resolving sequence read and write errors, (2) improved time and space complexity than existing fountain-codes based approaches, and (3) an encoding approach to handle constraints on structure (such as, palindrome and non-local repeats) and biological function (for example, elimination of start/stop codons) that other encoding approaches do not consider.

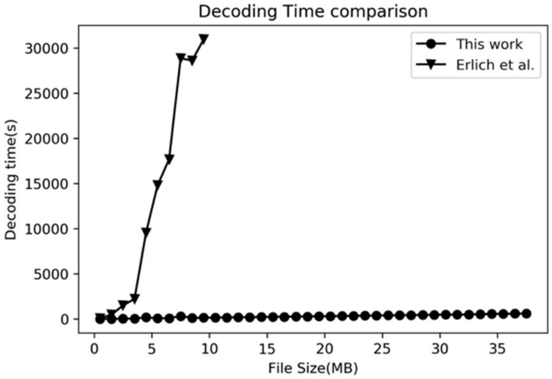


Fig 3. Decoding time of the original fountain code algorithm compared to our in-house decoding algorithm.

Our seqNAM encoding/decoding algorithms were analyzed using memory and runtime profiling tools to improve space and time efficiency, which would decrease both memory usage and running time of our algorithms. Following these improvements, simulations were run to test the robustness of our encoding algorithm and to test how much error the algorithm can handle for different random binary files ranging from 512KB to 35 MB in size. For comparison, **Figure 3** shows the decode time of Erlich's algorithm[1] — which is from the original paper introducing the use of fountain codes for DNA-based information storage — compared to our custom algorithm. Because of the long decoding time of Erlich's algorithm, we were unable to perform as many tests as were performed for our algorithm. Our algorithm exhibits a linear growth in decoding time while Erlich's shows an exponential growth curve.

In addition to evaluating decoding t ime, we performed degradation tests on the algorithms (**Figs 4a and 4b**). Errors were introduced into encoding files by randomly choosing (with replacement)

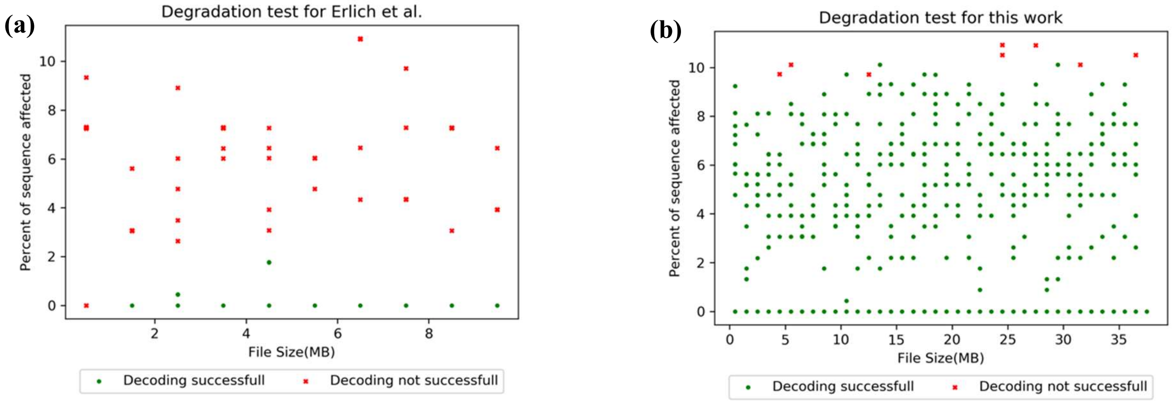


Fig 4. Degradation tests as a function of file size using (a) Erlich's algorithm and (b) our custom algorithm.

[1] Erlich, Y, and Zielinski, D. "DNA Fountain enables a robust and efficient storage architecture." *Science* 355, no. 6328 (2017) 950-954.

a DNA sequence encoding part of the file and inserting an error in a random position of the sequence. Error frequencies were varied so that the final percent of affected sequences ranged between 0 and ~11%. The errors introduced into the file were either insertion, deletion, or mutation related. In the insertion error, a randomly selected nucleotide is inserted in a random location in the sequence. In the deletion error, a random nucleotide is removed, and for a mutation error a random nucleotide is replaced. All the files that were tested were randomly generated binary files.

We are in the process of validating the algorithm performance under real-world conditions. For this, we used our encoding algorithm on a JPEG image of 13 KB size. The image is shown in **Figure 5**. The sequences output from our encoding algorithm were ordered from Integrated DNA Technologies, which shipped them directly to another DNA technology company, GENEWIZ, for sequencing. The sequences are currently being processed by GENEWIZ.

Fig 5. File encoded via our custom algorithm.

We have also devised an improved encoding process for seqNAM based on the shortest-path problem of a directional weighted graph (**Fig 6**). As a reminder, seqNAM encodes information directly into DNA sequences. In our approach, the graph is defined by the incoming bit-stream that is to be encoded. Each vertex corresponds to an oligo that encodes a specific substring of the bitstream. A directed edge (p,q,w) represents the encoding step going from p to q with weight w, where w (w > 0) represents the cost of encoding the next part of the bit-stream as q after p. Therefore, for a given input, each possible path through the graph from a source node to a sink node (source and sink node are non-coding) is a possible encoding event. The weight/cost on each edge in the graph comes from various pre-calculated constraints. Therefore, the best encoding for a given input is the shortest possible path from source to sink. If there is no path from source to sink, then there is no valid encoding for that input. Our algorithm requires these key components: (1) a dictionary compiling all of the information for each valid oligonucleotide that can be used for encoding; (2) a bit-string to oligonucleotide map; (3) an encoder to convert an input bitstream into a graph; and (4) a validation step to ensures that the output of the encoder is valid.

The dictionary, map, and encoder are static elements, meaning that they will not change and are not dependent on the data to be encoded. The graph is created by the encoder, which takes into

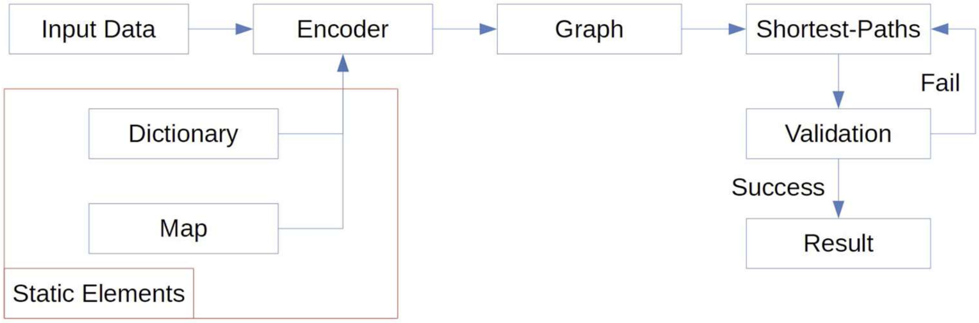


Fig 6. Schematic of the encoding process.

account static, local constraints in order to determine edge weights. The list of shortest paths are determined from the graph and passed to the final validation step. The validation step examines potential paths for violation of non-local constraints such as repeat and/or palindrome sequences and any other structural or functional issues that would violate predetermined constraints, and determines the shortest path sequence that passes all tests. The validation step is the most difficult, as it requires examination of non-trivial constraints. For example, we wish to remove sequences with undesirable binding or secondary structures. As a means to speed up determination of secondary structure, we have been examining the use of Long Short-term Memory neural networks, Convolution neural networks, and transformer networks.

Similar to that used by Singh *et al.* for RNA structure prediction (SPOT-RNA), we designed a LSTM-based neural network architecture to determine secondary structure and binding parameters for DNA. Unlike SPOT-RNA, which used base-pair binding energy as input, the input to our NN was the DNA sequence. Training data was generated using NUPACK to determine secondary structure of randomly generated sequences. The input to the NN was a one-hot encoded DNA sequence, and the target output was a one-hot encoded structure sequence. We tested LSTM RNNs, Convolution Neural Networks (CNNs), and an architecture combining both. As can be seen in **Figure 7**, the CNN and LSTM architectures individually perform about the same, with test set accuracies around 70%, while the combination of the two architectures achieved over 75% accuracy. These accuracies are adequate, but our current work is looking to both improve this result as well as speed up processing.

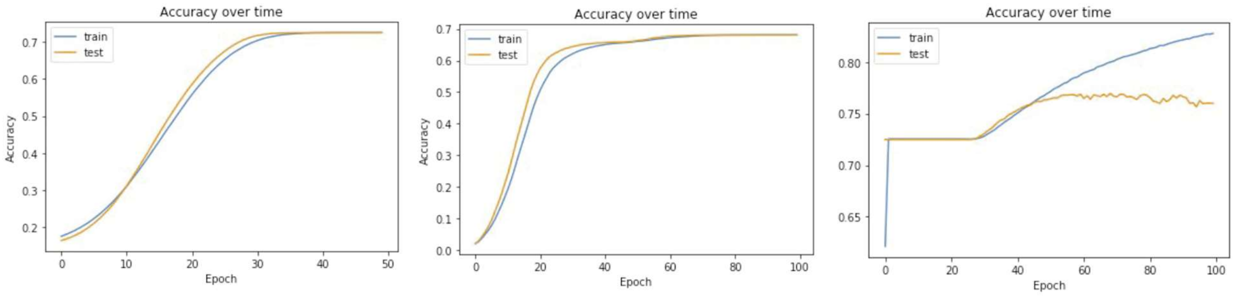


Fig 7. Full structure prediction using an LSTM, CNN, and combined LSTM-CNN.

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 publicly available and Dr. Mike Tobiason, who created the software during his PhD, has returned to Boise State as a postdoctoral fellow and uses the software to design and select DNA sequences for *Tasks 3, 4, and 5*. We have recently purchased high-performing computational resources to 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.

Building on our scaffold design work highlighted in the Year 2 annual report, we identified DNA oligonucleotide "staple strands" that self-assemble with the scaffold as a potential source of noise during imaging with DNA-PAINT. Our origami requires over 100 individually synthesized staple strands creating a challenge for quality control due to variability during synthesis. In response, we are developing a defect analysis process for DNA using mass-spectrometry. Mass spectrometry has several advantages over common gel electrophoresis approaches: (1) it is cost-effective, (2) exhibits improved throughput compared to gel based approaches, and (3) can potentially 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. As such, we explored the experimental parameter phase space of matrix-assisted laser desorption/ionization (MALDI) mass spectrometry. Specifically, our team evaluated combinations of two different matrices, three different sample application methods, two different types of MALDI plates, and a plethora of different MALDI instrument (mainly laser) settings. We found that the best conditions are not sufficient for our desired mass resolutions and plan on pursuing higher resolution electrospray ionization mass spectrometry approaches during the next reporting period.

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

The two major efforts during this reporting period have been associated with the introduction of editing to our read/write protocol for dNAM, as well as an assessment of imager length on SRM performance. The ability to edit data written into our DNA memory platform has the potential to take dNAM from an archival storage solution to a responsive information storage system. We aim to edit single data sites directly on dNAM during a single reading experiment. Further, short imager lengths are a prerequisite for 3DNAM. Thus, we are working toward defining imager lengths that lead to optimum SRM performance. Doing so will enable higher data density using the 3DNAM platform.

4.1 dNAM editing. We based our data editing approach on the DNA strand-displacement technique as it does not involve chemical modification to displace DNA oligonucleotide strands. In response to the intrinsic challenges working with a dynamic molecular system, we designed two editing protocols. Our initial editing approach was designed to have two distinct editing strands — one for denoting a "0" and the other for "1". While this approach was not successful due to inconsistent imager strand hybridization, the knowledge gained from these experiments provided a roadmap for finely tuning reaction parameters. Based on this insight, we designed our second editing approach to ensure docking sites are always available for imager hybridization. In this approach, the imager strands are included with a pool of site specific "blocker" and "eraser" strands. We determine that the image quality is not sufficiently high for a robust implementation of this technique. However, from these initial experiments, we have developed strategies to overcome inconsistent site to site availability, such as increasing the editing strand stability.

4.2 dNAM imager performance optimization. We are building a comprehensive library of imager lengths toward achieving the best possible SRM horizontal and axial resolution. This supports DNA-PAINT experiments in general, as well as providing a solid foundation for identifying the shortest possible imager probes (and therefore, the highest data density) for 3DNAM. The parameters under investigation include nucleotide sequence, nucleotide length, oligonucleotide secondary structure, and chemical modifications that can bring additional stability to the pairing. First, we tested the shortest imager strand reported in literature (7 nts long respect to our 10 nts previously used) on our dNAM system (**Fig 8**). Building off this successful demonstration of short imager strands, we inserted a small DNA mini-hairpin into the imager strand. The mini-hairpin's secondary structure enhanced imager probe hybridization, which resulted in improved DNA-PAINT image resolution (**Fig 9**).

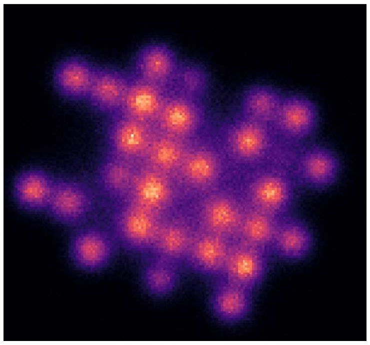


Fig 8. Representative image of dNAM platform with the 7 nts imager probes. The new sequence has been validated for future experiments.

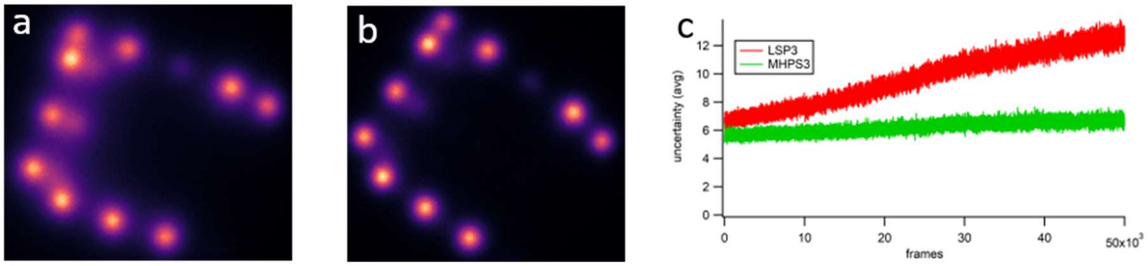


Fig 9. Representative images demonstrating dNAM image quality when using (a) linear or (b) mini-hairpin imager probes. From the averaged images, the mini-hairpin imager probes result in a well-defined pattern. (c) Uncertainty values associated with fitting the point spread function at each site for linear (red curve) and mini-hairpin (green curve) imager probes confirms the source of the improved quality. The experiment has been repeated 4 times with the same outcome.

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

5.1 Sub-nanometer imaging resolution for SRM. Noise, in the form of vibration and drift, is a huge factor in image resolution. While we have demonstrated an active drift correction system that enables sub-nanometer RMS stability during lengthy, many-frame super-resolution imaging experiments (see Year 2 Annual Report), we expect additional resolution gains with our recently procured environmental chamber and our custom super-resolution microscope that is under development (**Fig 10**). The environmental chamber will provide constant temperature — and thus reduce drift — during imaging. Further, the custom super-resolution microscope is designed to maximize stability, minimize the transmission of low frequency vibrations, and reduce drift. Specifically, it will consist as one monolithic mechanical system with no macroscopic moving parts. In doing so, major sources of non-statistical noise from the experimental set-up and environmental conditions will be minimized. This will allow us further to approach theoretical ~2nm limit on resolution with our current technique and materials.

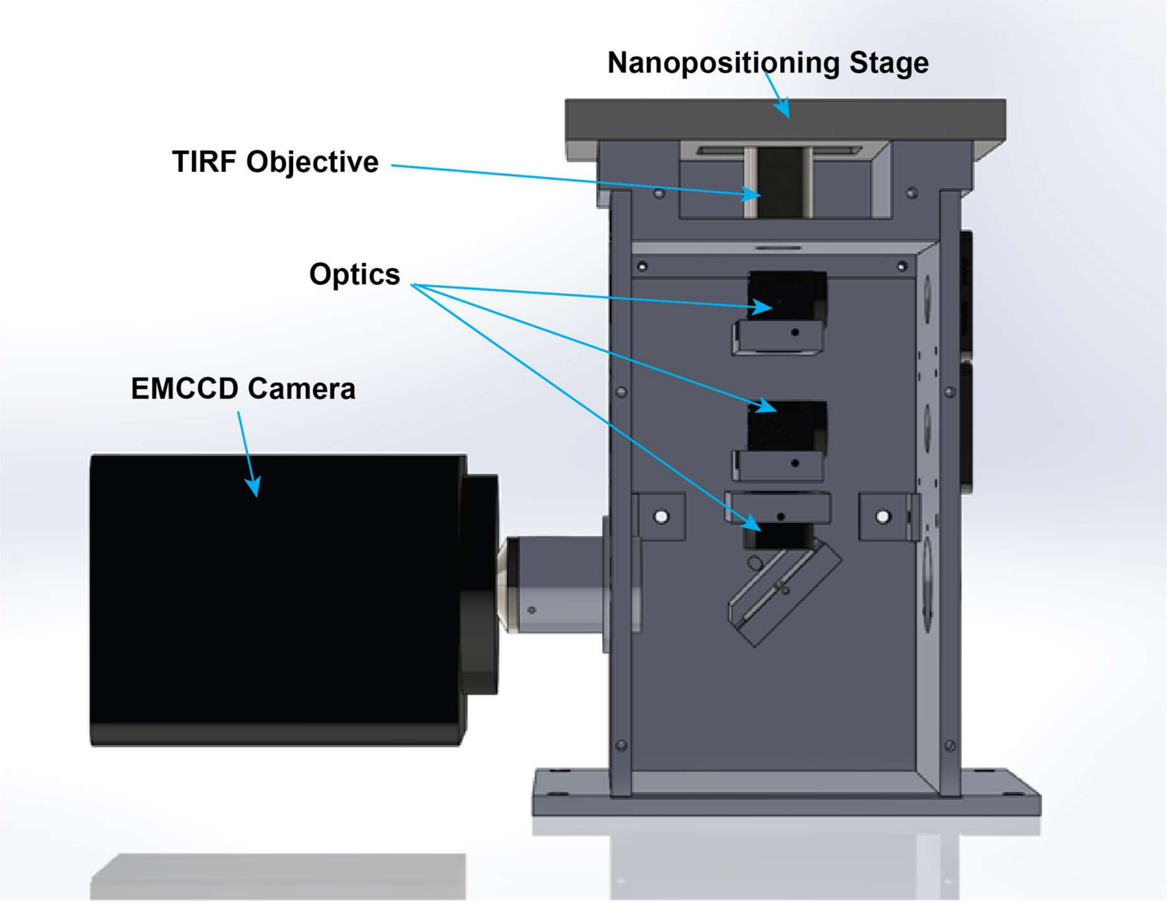


Fig 10. Solid part assembly design for custom super-resolution microscope body, with commercially available microscopy components (labeled) attached.

5.2 Time-Correlated Super-Resolution Microscopy (TCSRM) as an Enabling Technology Toward Three-Dimensional Nucleic Acid Memory (3DNAM). Super-resolution techniques developed in recent years have revolutionized biological and biomedical research, allowing optical imaging well below the classic diffraction limit of light. The significance of this imaging technique is highlighted by the 2014 Nobel Prize in Chemistry "*for the development of super-resolved fluorescence microscopy*". Super-resolution is achieved by controlling the state of fluorophores such that only a small subset of them are detectable at any given time. DNA points accumulation for imaging in nanoscale topography (DNA-PAINT) is the only super-resolution imaging method that can achieve sub-5 nm spatial resolution. Although DNA-PAINT has proven to be extremely precise in its ability to localize fluorescent probes in the lateral dimensions, achieving similar resolution in the axial dimension is difficult. A number of point-spread function engineering approaches, such as 3D stochastic optical reconstruction microscopy (3D-STORM), biplane photoactivated localization microscopy (PALM), and interferometric PALM have managed a modest improvement. However, the resolution in the axial direction is still a factor of five worse than in-plane resolution. Thus, in pursuit of 3DNAM, we are collaborating with Ben Johnson—an Assistant Professor within the Electrical and Computer Engineering Department. Leveraging his experience with integrated circuit design, CMOS imagers, and SPADs, he and his PhD student, Mehdi Sharsan are heading up the effort in the design and development of time-correlated imaging arrays for 3DNAM. Once developed, TCSRM will represent a one-of-a-kind microscope, positioning Boise State as an international leader in super-resolution microscopy with a power tool enabling the probing of biological structure in three dimensions with high lateral and axial resolution.

To enable 3DNAM, we are developing a custom imaging array that combines high resolution, high light sensitivity, and high timing sensitivity. Conventional imagers have a lengthy exposure time to capture an image in a low-light environment, as we have in imaging the fluorescence of 3DNAM. Due to this exposure time, conventional imagers are incapable of extracting fluorescent lifetime. Our time correlated imager (TCI) under development uses single-photon avalanche diodes (SPAD) which have a binary response to a single photon, meaning we can extract the exact moment of a photon's arrival. There are commercially available SPAD imagers; however, they are unsuitable for our application, as they either have only a single pixel or have poor photon detection efficiency. Fortunately, SPADs can be integrated directly into integrated circuit technology, meaning we can develop our own imager with supporting circuits to extract lifetime information. **Figure 11A** shows the cross section of the SPAD we will use to implement our TCI. **Figure 11B** shows a simplified architecture of our first TCI prototype. It is comprised of a 16x16 SPAD array, column-level monostable circuits to stretch SPAD events and then reset the SPAD, and shared time-to-digital converters (TDC) that convert the photon's arrival time with respect to the laser into a digital code. Our TDC has a timing resolution of 62 ps and a selectable range to accommodate up to an 80MHz laser pulse repetition rate. Our next prototype will include lifetime computation directly on chip to compress data.

We are developing the TCI in a 180nm semiconductor process provided by a commercial foundry (X-FAB). We also use an industry standard toolset (Cadence) to simulate and verify the design of our imager. Upon design completion in January 2021, we will submit the design to the foundry for fabrication. Note that the time correlated imager we are creating has potential for commercial development. Single photon detection and precision timing capabilities are only available as bulky and low-throughput devices. Thus, by developing TCI, we are not only providing an enabling technology for 3DNAM, but are also addressing an unmet commercial need for this class of scientific instrumentation.

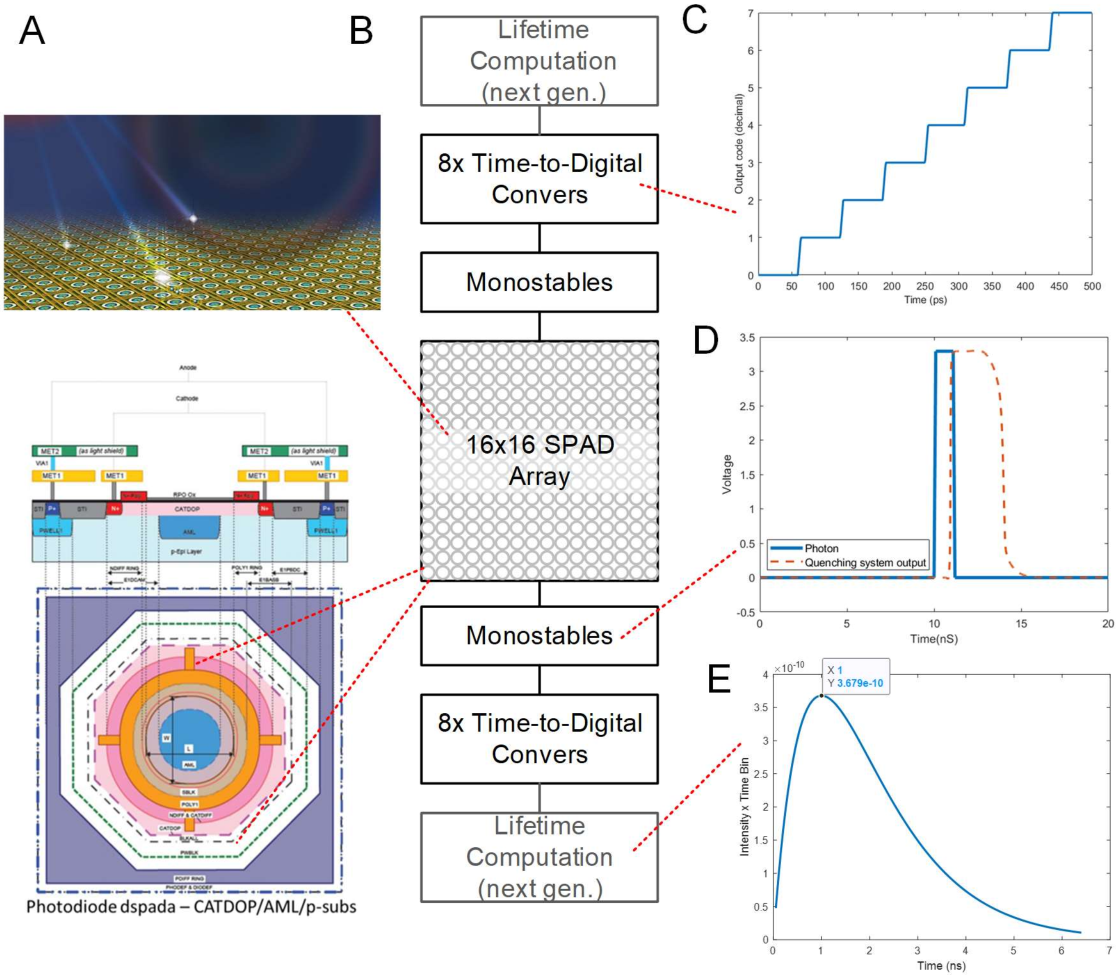


Fig 11. (A) SPAD cross-section design for high photon detection efficiency. (B) Architecture of time correlated imaging array. (C) Simulated linearity of the time-to-digital converters. (D) Example of monostable output due to a single photon. (E) Example of fluorescent lifetime extraction using the center-of-mass method.

5.3 Three-Dimensional Nucleic Acid Memory (3DNAM) Architecture. To maximize information density in the axial direction in 3DNAM (**Fig 12**), imager strands must be short. To minimize their crosstalk, their sequences must be orthogonal. And to maximize the probability that they find their target, both the imager strand and its respective binding site must be *mutually available* to interact. As the imager length decreases, so does its discrimination power to resolve a binding site among a population of similar but not equal targets. In addition, the sequences of imager and data strands directly affects their ensemble of secondary structures in solution and their mutual availability; both of which sway their kinetic uniformity.

In support of designing a DNA origami system capable of storing information in three dimensions, we have simulated an array of architectures and their response when probed using TCSRM. 3DNAM

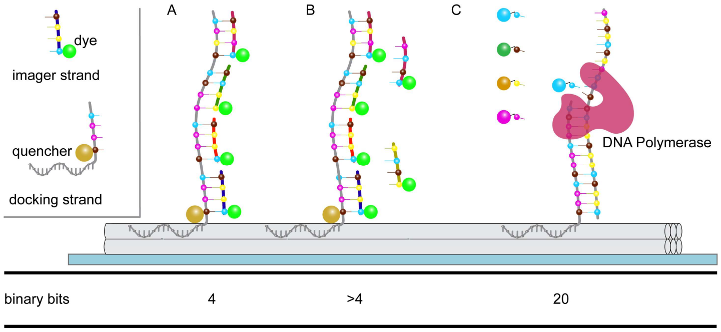


Fig 12. Schematic of 3DNAM architecture. Data strands extend from the origami substrate (grey cylinders). Fluorescence lifetime is dependent on the distance between the quencher (gold) and dye molecules (green). Data density increases from (a-c) with shorter imagers (colored strands) and additional fluorescence color-channels. For example, imager strands of 4 nt will read a 16 nt docking site with 4 or 8 bits/site if one or four-color channels are respectively deployed (A). The imager strands are intentionally overlapping in (B). Upgrading the docking site with a double-stranded primer in (C) and functionalizing a polymerase (plum) with a quencher allows 3DNAM to be resolved at the single-nucleotide level via time-resolved sequencing.

involves energy transfer between dye-labeled imager strands that alters their fluorescence lifetime depending on the distance between the acceptor and the donor. Theoretical measurements of the fluorescence lifetime using an imaging timing detector show sub-nanometer resolution for acceptor-donor distance (**Fig 13**); in line with past experiments carried out with other sensor technologies.

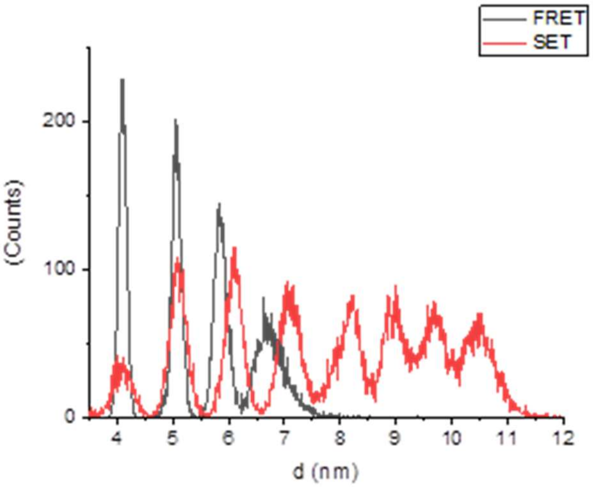


Fig 13. Results of 3DSRM simulation of multiple emittters spaced 1 nm apart along vertical axis. Simulation shown using two different energy transfer processes, Forster energy transfer (FRET) and surface energy transfer (SET). Simulation shows ability to resolve emitter locations with sub-nanometer precision.

IV. Demonstration of economic development and impact

Demonstration of Economic Development and Impact	**Year 1 Reporting Period 07/01/2018–06/30/2019**	**Year 2 Reporting Period 07/01/2019–12/31/2020**	**Current Reporting Period 07/01/2020–01/01/2021**
External Funding	$ 1,549,995	0	0
Number of External Grants	3	1 submitted, not recommended for funded	0
Private Sector Engagement	14 companies	2 companies, 1 VC group	1 VC group
University Engagement	11 universities	~20 universities	~5 universities
Federal Agency Engagement	5 agencies	4 (NSF, SRC, NRL, NIST, IARPA)	3 (NSF, NIST, IARPA)
Industry Involvement	2 companies	2 companies (Micron, SRC)	2 companies. (Micron, SRC)
Patents	0	1 Patent Application	1 Patent Application
Copyrights	0	0	0
Plant Variety Protection Certificates	0	0	0
Technology Licenses Signed	0	0	0
News Releases	3 articles	0	0
Start-up Businesses Started	0	1	0
Jobs Created outside of Boise State	0	6	0

External Funding

During the previous reporting period, we pursued the $1.5M National Science Foundation 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 both high risk and high reward. The benefit of applying is that our team is better positioned to resubmit the proposal next year and our team is moving forward on gathering the preliminary results needed to derisk the investment from NSF. The NAM Institute is also targeting the NSF Partnerships for Innovation (PFI) Grant mechanism to help build our entrepreneurial network and knowhow when exploring spinning off a company related to this project. In addition, the Institute is targeting the NSF Engineering Research Center (ERC) Planning Grant mechanism to establish the scholarly network and knowhow to create a more sustainable foundation for the Institute to support future faculty, staff, and students to translate their ideas into the marketplace.

Engagement

Semiconductor Research Corporation. Will Hughes facilitated a series of communications and conversations with the President and Chief Scientist of the Semiconductor Research Corporation (SRC) for the Micron School of Materials Science & Engineering and specifically the Nucleic Acid Memory Institute to become an SRC-designated research center. As shared, becoming an SRC-site would mean direct investment from the SRC and its industrial consortium to build-out NAM technologies and more broadly work at the interface of semiconductor fabrication and synthetic biology. While the conversations were productive, they have been put on hold until this summer because of the financial volatility related to COVID-19.

National Institute of Standards and Technology. The NAM Institute and specifically Drs. Luca Piantanida and Will Hughes have been invited to create a DNA nanotechnology tutorial for publication with Dr. Alex Little, who is the Director of the Microsystems and Nanotechnology Division at the National Institute of Standards and Technology (NIST). The tutorial is intended to help translate the scientific findings in their labs to become engineering principles for the DNA nanotechnology community. Our contribution to the work is focused on our findings during this project and we are hopeful that this will translate to additional collaborations.

Intelligence Advanced Research Projects Activity Program. The NAM Institute submitted its first major manuscript to Nature last year, which is still under review. The manuscript was independently reviewed by Dr. David Markowitz, who is the Program Manager for the Molecular Program Storage program at the Intelligence Advanced Research Projects Activity Program (IARPA). IARPA is an organization within the Office of the Director of National Intelligence responsible for leading research to overcome difficult challenges relevant to the United States Intelligence Community and Dr. Markowitz is an important figure in the DNA memory community because he is shaping its research agenda and supporting its financing. According to Dr. Markowitz and his team, our NAM work is "very innovative and we think it has significant potential as a next-generation storage medium." In response, Will Hughes has been invited to help facilitate an IARPA workshop on Biologically-Templated Transistors; which a focus on helping IAPRA create its request for proposals (RFP) that build on the work of the NAM Institute.

Business Development

Steven Burden, who successfully completed his PhD in Biomolecular Sciences, graduated December 2019 as a member of the NAM Institute. 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. In all, FACible BioDiagnostics employs 6 people — three full time and three part time.

The NAM Institute is considering spinning-off a 2nd company related to our research. We have a ***provisional patent*** for digital Nucleic Acid Memory (dNAM), which was filed on 7/24/2020 and given serial number 62/705,995. The utility application is currently being filed at Boise State and we have until July 2021 to complete the application process. Toward this effort, we are targeting the NSF Partnerships for Innovation (PFI) funding opportunity as a vehicle to translate emerging technologies from proof-of-concepts to commercially viable products.

V. Numbers of student, staff, and faculty participation

Classification	**Number**
Tenured or Tenure Track Faculty	4 (*2 full professors, 2 associate professors*)
Research Faculty	1 (*started a tenure-track faculty position*)
Project Manager	1 (*also focused on business development*)
Senior Lab Research Associate	1 (*manages the laboratory & supports team*)
Postdoctoral Fellows	4 (*performing at a research faculty level*)
Graduate Students	7 (*3 of the 7 have graduated*)
Undergraduate Students	9 (*4 female and 5 male*)

New Personnel

The no-cost extension provided by IGEM-HERC for Year 2 funding enabled the team to hire Mike Tobiason in support of designing, synthesizing, and characterizing NAM technologies. Tobiason was a previous PhD student in the NAM Institute having earned his PhD on his work titled, "*Engineering Kinetically Uniform DNA Devices*." With Tobiason's DNA biotechnology expertise along with his knowledge of NAM, he has been able to make an immediate impact to the project. This is particularly important because not everyone on the team is able to work in the lab because of COVID.

In addition to the postdoctoral research scientists, we also hired a new PhD student, Sarah Kobernat, in support of designing, producing, and optimizing large DNA origami scaffolds (*Task 3*). Sarah is supporting the Vertically Integrated Project through mentoring undergraduate students on scaffold design. And as described in *Task 5.2* and *5.3*, we are collaborating with Benjamin Johnson—an Assistant Professor within the Electrical and Computer Engineering Department—in the design and implementation of TCSRM for 3DNAM. We are providing direct support through IGEM-HERC for his PhD student, Mehdi Sharsan, who is designing the TCI array for 3DNAM.

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. During the Fall 2020, due to COVID related lab personnel restrictions, VIP students focused their efforts on literature reviews in the fields of DNA-based memory, DNA-PAINT and DNA Nanotechnology. As these fields are rapidly advancing, the literature reviews were intended to ensure we are in lockstep with the latest breakthroughs. Students were tasked with searching databases of scientific journals to identify high priority research articles, read and discuss the importance of these articles, and summarize their findings through an internal newsletter. In the future, it is expected that the summaries will be used to write traditional review papers as well as a resource for our bi-weekly research group meetings.

VI. Description of future plans

Team Management – Integration and graduation

- Continue to manage the financial and health risks of COVID-19; including but not limited to budget cuts at Boise State and our reduced experimental lab activity.
- Pursue and submit the next round of grant opportunities. Target opportunities include the NSF Partnerships for Innovation Program, Engineering Research Center Planning Grant, and the NSF SemiSynBio-III.
- Continue to help the postdoctoral fellows identify the intellectual space they want to lead; meeting with them to establish their professional development plans. As of today, one is seeking a faculty position and two have shown interest in starting a company related to our work.
- Rekindle collaborations with key internationally recognized research groups; with an eye towards cross-training our labs. Targeted groups include Paul Rothemund at Caltech, Alex Little at NIST, and George Church at Harvard.

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

- Experimentally validate the seqNAM encoding/decoding algorithms and publish results.
- Expand the training data for the dNAM system and train a neural network to optimize image resolution and decoding.
- To increase image resolution, perform noise cancellation using GAN or Autoencoder.
- Improve the error correction algorithm so that we can recover the file using a minimal amount of error correction bits.

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.

- Synthesize double-stranded DNA plasmids for larger scaffold production.
- Produce and assess large single-stranded DNA scaffolds from phage cultured *E. coli.* Quantify yield and optimize or scale up as needed.
- Synthesize origami using larger scaffolds and then evaluate the origami by electrophoresis, AFM and SRM. Compare error rates (false positives and negatives) to smaller origami.

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

- In support of advancing our dNAM platform, continue optimizing our read and write approach with the inclusion of editing.
- Continue investigating optimal imager strand parameters, including new nucleotides sequences, different lengths, and secondary structures.
- Explore the application of blocked nucleic acid (BNA) and other DNA analogues as imager probes and editing strands. We hypothesize that BNA can increase the stability of the dNAM platform, which will increase resolution quality of DNA-PAINT images.
- Combine knowledge and technology advances of our dNAM research to evaluate 3DNAM (which is a derivative to seqNAM) as a viable system to increase information density in NAM.

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

- Manufacture, assemble, and test custom SRM system.
- Improve active drift correction algorithm and user interface to enable routine use by other users and integration into future NAM experiments.

- Experimentally test 3DSRM to determine limits of resolution. Initial experiments will combine separate imaging sensor and timing sensor using a beam splitter, allowing characterization of technique capabilities before fabrication of imaging timing array sensor.

VII. Summary of Budget Expenditures

The below table summarizes expenditures associated with the project from July 1, 2020 to December 30, 2020. It includes a budget adjustment associated with the $4K reduction (which we applied to *Travel*) as part of the 5% statewide budget hold-back. *Salary* and *Fringe* supported our graduate students and postdoctoral research scientists. *Other Expenses* were used to purchase supplies to process modified and unmodified DNA oligos into dNAM, super-resolution microscopy supplies, productivity software, sequencing of DNA structures to validate seqNAM encoding/decoding algorithms. Major *Capital* purchases include a Dell PowerEdge DSS 8440 system to support the development and training of ANN-based algorithms for more efficient image analysis, a high-precision optical table to further minimize noise being captured in SRM images, and a qPCR system to provide rapid design, build and test cycles for DNA origami.

Category	Original Year 3 Budget	Year 3 After State Hold-back	Expenditures	Encumbrances	Remaining Budget
Salary	$288,720.00	$288,720.00	$103,136.66	--	$185,583.34
Fringe Benefits	$99,311.00	$99,311.00	$30,128.17	--	$69,182.83
Other Expenses	$123,066.00	$123,066.00	$9188.59	$3,080	$110,797.41
Travel	$25,000.00	$21,000.00	--	--	$21,000.00
Capital	$100,000.00	$100,000.00	--	$91,866.70	$8,133.30
Student Costs	$30,403.00	$30,403.00	$19,276.00	--	$11,127.00
Total	$666,500.00	$662,500.00	$161,729.42	$94,946.70	$405,823.88

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	5
Number of VIP Students Enrolled (grad and undergrad)	9~10
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.