SCIENTIFIC REPORTS

OPEN

Demonstration of End-to-End Automation of DNA Data Storage

Christopher N. Takahashi[1], Bichlien H. Nguyen[1,2], Karin Strauss[1,2] & Luis Ceze[1]

Received: 26 October 2018
Accepted: 5 March 2019
Published online: 21 March 2019

Synthetic DNA has emerged as a novel substrate to encode computer data with the potential to be orders of magnitude denser than contemporary cutting edge techniques. However, even with the help of automated synthesis and sequencing devices, many intermediate steps still require expert laboratory technicians to execute. We have developed an automated end-to-end DNA data storage device to explore the challenges of automation within the constraints of this unique application. Our device encodes data into a DNA sequence, which is then written to a DNA oligonucleotide using a custom DNA synthesizer, pooled for liquid storage, and read using a nanopore sequencer and a novel, minimal preparation protocol. We demonstrate an automated 5-byte write, store, and read cycle with a modular design enabling expansion as new technology becomes available.

Storing information in DNA is an emerging technology with considerable potential to be the next generation storage medium of choice. Recent advances have shown storage capacity grow from hundreds of kilobytes to megabytes to hundreds of megabytes[1–3]. Although contemporary approaches are book-ended with mostly automated synthesis[4] and sequencing technologies (e.g., column synthesis, array synthesis, Illumina, nanopore, etc.), significant intermediate steps remain largely manual[1–3,5]. Without complete automation in the write to store to read cycle of data storage in DNA, it is unlikely to become a viable option for applications other than extremely seldom read archival.

To demonstrate the practicality of integrating fluidics, electronics and infrastructure, and explore the challenges of full DNA storage automation, we developed the first full end-to-end automated DNA storage device. Our device is intended to act as a proof-of-concept that provides a foundation for continuous improvements, and as a first application of modules that can be used in future molecular computing research. As such, we adhered to specific design principles for the implementation: (1) maximize modularity for the sake of replication and reuse, and (2) reduce system complexity to balance cost and labor input required to setup and run the device modules.

Our resulting system has three core components that accomplish the write and read operations (Fig. 1a): an encode/decode software module, a DNA synthesis module, and a DNA preparation and sequencing module (Fig. 1b,c). It has a bench-top footprint and costs approximately $10 k USD, though careful calibration and elimination of costly sensors and actuators could reduce its cost to approximately $3 k–4 k USD at low volumes.

Before a file can be written to DNA, its data must first be translated from 1's and 0's to A's, C's, T's, and G's. The encode software module is responsible for this translation and the addition of error correction into the payload sequence (see the Methods section and work by Richard Hamming[6]). Once the payload sequence is generated, additional bases are added to ensure its primary and secondary structure is compatible with the read process and the DNA sequence is sent to the synthesis module for instantiation into physical DNA molecules.

The DNA synthesis module is built around two valved manifolds that separately deliver hydrous and anhydrous reagents to the synthesis column. Our initial designs used standard valves, but the dead volume at junction points caused unacceptable contamination between cycles. Therefore, we switched to zero dead volume valves[7]. The combined flow path is then monitored by a flow sensor, whose output is coupled to a standard fitting; the fitting can be coupled to arbitrary devices, such as a flow cell for array synthesis[8] or, in this case, adapted to fit a standard synthesis column. Once synthesis is complete, the synthesized DNA is eluted into a storage vessel, where it is stored until retrieval.

When a read operation is requested, the stored DNA pool's volume is reduced to about 2 μL to 4 μL by discarding excess DNA through the waste port. A syringe pump in the DNA preparation and sequencing module then dispenses our single-step preparation/sequencing mix (Fig. 1d) into the storage vessel; positive pressure pushes the mixture into the ONT MinION's priming port (Figs 1b,c). We chose the MinION as our sequencing device due to its low cost, ease of automation, and high throughput. However, it is neither capable of reading unmodified DNA, nor is it optimized for reading short DNA oligonucleotides[9]. In particular, we have observed that reads shorter than 750–1000 bases tend to get missed or discarded by the MinION's software. To mitigate these limitations, we developed a single-step MinION

[1]School of Computer Science and Engineering, University of Washington, Seattle, Washington, USA. [2]Microsoft Research, Redmond, Washington, USA. Correspondence and requests for materials should be addressed to C.N.T. (email: cnt@cs.washington.edu)

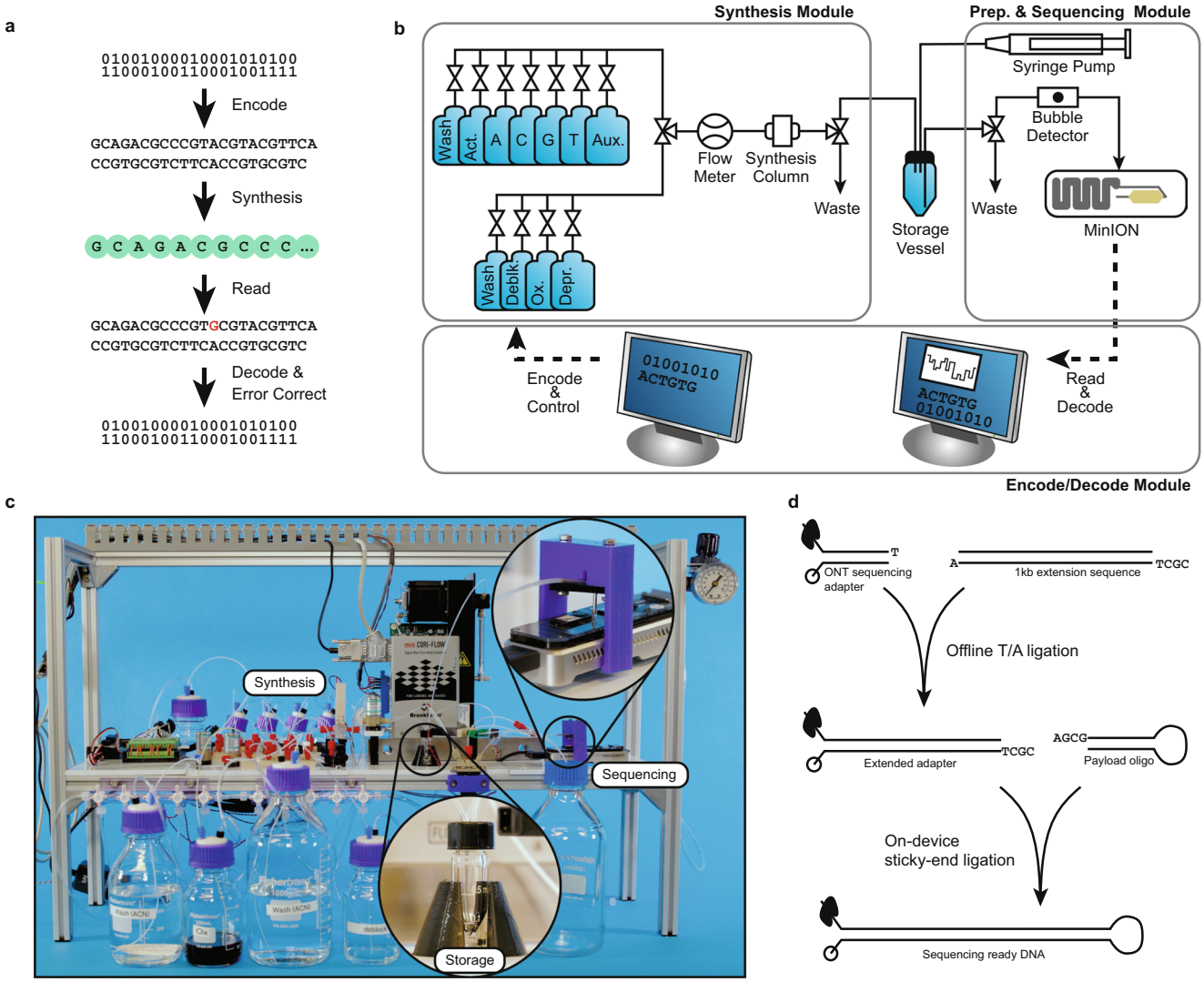

Figure 1. An overview of the write-store-read process. Data is encoded, with error correction, into DNA bases, which are synthesized into physical DNA molecules and stored. When a user wishes to read the data, the stored DNA is read by a DNA sequencer into bases and the decoding software corrects any errors retrieving the original data. (**a**) The logical flow from bits to bases to DNA and back. (**b**) A block diagram representation of the system hardware's three modules: synthesis, storage, and sequencing. (**c**) A photograph showing the completed system. Highlighted are the storage vessel and the nanopore loading fixture. The majority of the remaining hardware is responsible for synthesis. (**d**) Overview of enzymatic preparation for DNA sequencing. An arbitrary 1 kilobase "extension segment" of DNA is PCR-amplified with TAQ polymerase, and a Bsa-I restriction site is added by the primer, leaving an A-tail and a TCGC sticky end after digestion. The extension segment is then T/A ligated to the standard Oxford Nanopore Technology (ONT) LSK-108 kit sequencing adapter, creating the "extended adapter," which ensures that sufficient bases are read for successful base calling. For sequencing, the payload hairpin and extended adapter are ligated, forming a sequence-ready construct that does not require purification.

preparation protocol that requires only payload DNA and a master mix containing a customized adapter (Fig. 1d) with a 1 kbase extension region, T4 ligase, ATP, and a buffer. Each payload sequence is constructed to form a hairpin structure with a specific 5′ 4-base overhang. The customized adapter has a complementary overhang, which aids T4-mediated, sticky-ended ligation. To sequence, the payload and master mix are combined and incubated at room temperature for 30 minutes. Thereafter, the mixture is directly loaded into the MinION through the priming port. Since the introduction of air bubbles causes sequencing failure, we built a 3D printed bubble detector that valves off the loading port immediately after detecting the gas that is aspirated following the sample. This allows the system to load nearly the full sample without damaging the flow cell. Additionally, while not demonstrated here, other research suggests that random access via selective ligation over a small set of sequence identifiers (≈20) can be achieved using orthogonal sticky ends during preparation[10].

Once sequencing begins, the decode software module aligns each read to the 1 k base extension region and the poly-T hairpin. If the intervening region of DNA is the correct length, the decoder attempts to error check/correct the payload using a Hamming code with an additional parity bit; the code corrects all single-base errors and detects all

Figure 2. Synthesis and sequencing process quality. (**a**) Insertion, deletion, and substitution frequency by locus for a synthesized and PCR-amplified 100-mer. Below: An overview of errors. Above: An expanded view of the central 60 bases. The terminal 20 bases come from primers used in amplification and therefore are not representative of synthesis quality. (**b**) Combined write-to-read quality of synthesis, ligation, and sequencing. Bases −60 to −4 (below, grey) are adapter bases. Bases −3 to 0 (below, red) are the ligation scar. Bases 0 to 39 (below, blue) are the synthesized payload region with 8 bases of padding on the 3′ end. (**c**) Distribution of nanopore read lengths with unligated, 1D and 2D read lengths identified.

double-base errors. Once the payload is successfully decoded, it is considered correct if it matches a 6-base hash stored with the data. At this point, sequencing terminates, and the MinION flow cell may be washed and stored for later reuse.

Our system's write-to-read latency is approximately 21 h. The majority of this time is taken by synthesis, viz., approximately 305 s per base, or 8.4 h to synthesize a 99-mer payload and 12 h to cleave and deprotect the oligonucleotides at room temperature. After synthesis, preparation takes an additional 30 min, and nanopore reading and online decoding take 6 min.

Using this prototype system, we stored and subsequently retrieved the 5-byte message "HELLO" (01001000 01000101 01001100 01001100 01001111 in bits). Synthesis yielded approximately 1 mg of DNA, with approximately 4 μg ≈ 100 pmol retained for sequencing. Nanopore sequencing yielded 3469 reads, 1973 of which aligned to our adapter sequence. Of the aligned sequences, 30 had extractable payload regions. Of those, 1 was successfully decoded with a perfect payload. The remaining 29 payloads were rejected by the decoder for being irrecoverably corrupt.

Inspecting the sequencing data indicates that the low payload yield and decode rate was largely due to two factors. The first and primary factor is *low ligation efficiency*. Although chemical conditions should be optimal for T4 ligase, incomplete strands from the unpurified synthesis product likely out-competed full-length strands, leading to a poor apparent ligation rate of less than 10% (Fig. 2c). The second factor is *read and write fidelity*. To interrogate the write error rate, we synthesized a randomly generated 100-base oligonucleotide with distinct 5′ and 3′ primer sequences. The oligonucleotide was then PCR-amplified and sequenced with an Illumina NextSeq instrument to reveal: an error rate of almost zero insertions; <1% substitutions; and 1–2% deletions (Fig. 2a) for most positions, with increased deletions toward the 5′ end due to increased steric hindrance as strand length increases[11]. Literature suggests a nanopore error rate near 10%[9,12], so we also performed a synthesis-to-sequencing error rate analysis on an 89-mer hairpin sequence, encoding "HELLO" in its first 32 payload bases. Figure 2b shows the read error when aligned to the extended adapter and payload sequence. Bases −60 to −1 were directly PCR-amplified from the lambda genome and given a good baseline for nanopore sequencing fidelity under our conditions; bases 0 through +40 come from the payload region and characterize the total write-to-read error rate. The complex combination of these errors — especially deletions and read truncations — causes many strands to be discarded before a decoding attempt is made. Indeed, of 25,592 reads in this new dataset, 286 aligned well in the −100 to −1 region (score > 400) and contained enough bases to attempt decoding. Of those 251 had uncorrectable corruption, 11 had invalid checksum bases after correction, 8 were corrupted but correctable and of those 3 had hashes in agreement, 16 were perfect reads, and 0 were decoded but contained the wrong message.

We demonstrated the first fully automated end-to-end DNA data storage device. This device establishes a baseline from which new improvements may be made toward a device that eventually operates at a commercially viable scale and throughput. While 5 bytes in 21 hours is not yet commercially viable, there is precedent for many orders of magnitude improvement in data storage[13]. Infact, recent storage advances by Erlich *et al.*[2] of 2 Mbytes and Organick *et al.* of 200 Mbytes[3] demonstrate orders of magnitude improvements in the past two years and the underlying physics and chemistry show impressive upper bounds for density[3].

Furthermore, the modules and methods developed here are now being applied to other molecular computing projects internally. For example, by using a non-cleavable linker in the synthesis column and adding a reagent port for chip-synthesized DNA, we can use the same platform to perform a database query in DNA[14]. Additionally, our sequencing preparation protocol and loading hardware can be adapted for use with our digital microfluidics platform[15] and used as a readout for DNA strand displacement reactions.

Near-term improvements will focus primarily on system optimizations in synthesis, cycle count, and cost. Synthesis time can be reduced by 10–12 hours with the addition of heat in the cleave step[16]. Multiple writes (with or without reads) can be achieved by the addition of additional synthesis columns and a fluid multiplexer. Multiple reads can also be achieved with minor modifications (Supplemental Section 1) and exploiting the MinION flow cell's reusability. Additionally, a cost-optimized version could be designed by eliminating the syringe pump and flow sensor, both unnecessary if flow rates are well measured and calibrated. This could save approximately 60% of our current device's

Step	Volume (μL)	Time (s)
deblock	600	50
Act + {A, C, T, G} (1:1)	350	120
Act + Phos. reagent (1:1)*	350	900
Oxidizer	750	10

Table 1. DNA synthesis reagent parameters. *Only performed as final coupling step to add 5′ phosphate.

Reagent	Volume (μL)
Extended adapter	15
T4 DNA ligase (NEB: M0202)	5
DTT-free 10× T4 buffer*	20
ONT RBF	93
Nuclease-free water	64
Total	197

Table 2. Sequencing prep master mix. *DTT-free 1× T4 buffer: 50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM ATP.

cost at the expense of more laborious operation. Future improvements will focus on bringing storage density, coding, and sequencing yield up to parity with modern manual and semi-automated methods.

Methods

DNA synthesis. DNA synthesis was performed using standard phosphoramidite chemistry[17] without capping. Volumes and times, described in Table 1, used reagents purchased from Glen Research Corporation. For solid support (PN: ML1-3500-5), we used a BioAutomation 50 nmole scale synthesis column containing controlled porosity glass.

DNA cleavage was performed in 32% ammonia at room temperature for 1 hour before eluting. De-protection continued for an additional 11 hours in the same ammonia solution in the storage vessel.

Our system is fluidically configured as in Fig. 1b and electrically configured as in Supplemental Section 2.

Sequencing preparation. The extended adapter was constructed from a 1 kilobase fragment that was PCR-amplified from the lambda genome using hot start TAQ DNA polymerase (NEB M0496) with a Bsa-I restriction site added by the forward primer. The resulting fragment after digestion had a 3′ A overhang and a 5′-GCGT sticky end on the bottom strand. The fragment was then T/A ligated and prepped according to Oxford Nanopore Technology's (ONT) LSK-108 kit protocol, yielding the extended adapter with a four base sticky end.

The extended adapter was then mixed according to Table 2 into a sequencing master mix that is used in automated sequencing prep. Thirty minutes prior to sequencing, the master mix was combined with the hairpin oligo and incubated. DTT was left out of the T4 buffer because it damages the nanopores and causes sequencing to fail.

Nanopore sequencing. Nanopore sequencing was done with an Oxford Nanopore Technologies MinION using an MIN-107 R9.5 flowcell and MinKNOW 18.7.2.0 software. Base calling was performed in 4000 event batches using Albacore 2.3.1. The read length distribution and write-to-read quality test were loaded manually (as described in the instructions for LSK-108 sequencing kits); the end-to-end code, write, read, and decode experiment was loaded automatically from the storage vessel.

Coding and decoding. Prior to coding the user data ("HELLO" in ASCII bytes plus the hash consisting of the right most 12 bits of the SHA256 hash) was passed through a one time a one time pad to increase entropy similar to previous work[3]. One time pads

$$X_1 = (1\,3\,0\,1\,0\,1\,1\,0\,3\,2\,2\,2\,1\,1\,3\,1\,0\,2\,2\,2\,3\,2\,2\,2\,1\,1\,3\,2\,1\,3\,0\,0)$$

and

$$X_2 = (3\,1\,1\,2\,2\,1\,1\,2\,3\,0\,2\,1\,1\,0\,3\,2\,2\,0\,3\,3\,0\,2\,2\,0\,3\,3\,1\,0\,1\,3\,2\,2)$$

were used for the first and second experiment described in this paper respectively.

Data was coded using a two-layer scheme that stored 5 bytes over 32 dsDNA bases with an additional 13 bases of 3′ padding to compensate for lost fidelity near the read end (Fig. 2). The outer layer consisted of a (31, 26) Hamming code[6] over a four-symbol alphabet with a checksum base that detects all two-base read errors and corrects all single-base errors. The following equivalences were made for the sake of algebraic simplicity: A = 0, C = 1, G = 2, T = 3. We used modulo-4 arithmetic and the canonical generator matrix

$$G = (I \; - \; A^T),$$

along with the canonical parody check matrix

$$H = (A \; I),$$

where

$$A = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

and I is the identity matrix of the appropriate dimension. To increase error detection, 6 of the 26 data bases stored a 12-bit hash of the payload, which was checked after decoding to ensure data integrity. Source code is available in Supplemental Section 3.

For decoding, groups of 4000 reads were collected and base-called using ONT's Albacore software on 12 CPU cores. Reads that passed QC in Albacore were then aligned to the extended adapter and sequenced for further filtering. Only reads that appeared to have a correctly sized payload region between the adapter sequence and the poly-T hairpin were sent for error checking and decoding.

DNA alignment. All DNA alignment was done using the parasail parasail_aligner command line tool[18] with arguments -d -t 1 -O SSW -a sg_trace_striped_16 -o 8 -m NUC.4.4 -e 4. Alignments to the adapter sequence for decoding used the additional flag -c 20, while payload error analysis used flag -c 8.

References

1. Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in dna. *Science* **337**, 1628–1628 (2012).
2. Erlich, Y. & Zielinski, D. Dna fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
3. Organick, L. *et al*. Random access in large-scale dna data storage. *Nature Biotechnology* **36**, 242 (2018).
4. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nature Methods* **11**, 499–507 (2014).
5. Yazdi, S. M. H. T., Gabrys, R. & Milenkovic, O. Portable and error-free DNA-based data storage. *Scientific Reports* **7**, https://doi.org/10.1038/s41598-017-05188-1 (2017).
6. Hamming, R. W. Error-detecting and error-correcting codes. *Bell System Technical Journal* **29**(2), 147–160 (1950).
7. Hunkapiller, M. W. Zero dead volume valve *United States Patent #US4558845A* (1985).
8. Fodor, S. P. A. *et al*. Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767–773 (1991).
9. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* **17**, 239 (2016).
10. Potapov, V. *et al*. Comprehensive Profiling of Four Base Overhang Ligation Fidelity by T4 DNA Ligase and Application to DNA Assembly. *ACS Synthetic Biology* **7**(11), 2665–2674, https://doi.org/10.1021/acssynbio.8b00333 (2018).
11. LeProust, E. M. *et al*. Synthesis of high-quality libraries of long (150 mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Research* **38**, 2522–2540 (2010).
12. Jain, M. *et al*. MinION analysis and reference consortium: Phase 2 data release and analysis of r9.0 chemistry. *F1000 Research* **6**, 760 (2017).
13. Walter, C. Kryder's law. *Scientific American* **293**, 32–33 (2005).
14. Stewart, K. *et al*. A content-addressable dna database with learned sequence encodings. *Proceedings of the 24th International Conference On DNA Computing and Molecular Programming (DNA24)* **11145**, 55–70 (2008).
15. Willsey, M. *et al*. Puddle: A dynamic, error-correcting, full-stack microfluidics platform. In *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS' 19 (ACM, New York, NY, USA, 2019).
16. Glen Research. *The Glen Report: Deprotection Supplement*, https://www.glenresearch.com/reports/gr20-24 (2013).
17. Tanaka, T. & Letsinger, R. L. Syringe method for stepwise chemical synthesis of oligonucleotides. *Nucleic Acids Research* **10**, 3249–3260 (1982).
18. Daily, J. Parasail: SIMD c library for global, semi-global, and local pairwise sequence alignments. *BMC Bioinformatics* **17**, https://doi.org/10.1186/s12859-016-0930-z (2016).

Acknowledgements

This research was supported by a sponsored research agreement and gifts from Microsoft and DARPA under the Molecular Informatics Program (W911NF-18-2-0034).

Author Contributions

C.N.T. designed and built the hardware and software, performed all data analysis, and wrote the manuscript. C.N.T. and B.H.N. performed all experiments. B.H.N. advised on protocol development. K.S. and L.C. advised on all aspects. All authors read and edited the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-41228-8.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.