DNA Punch Cards: Encoding Data on Native DNA Sequences via Nicking

S Kasra Tabatabaei[1], Boya Wang[2*], Nagendra Bala Murali Athreya[3*], Behnam Enghiad[5], Alvaro Gonzalo Hernandez[4†], Jean-Pierre Leburton[3†], David Soloveichik[2†], Huimin Zhao[1,5,6 §], Olgica Milenkovic[3§]

[1] Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801, USA,

[2] Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, Texas, 78712, USA,

[3] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA,

[4] Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801, USA,

[5] Department of Chemical and Biomolecular engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801, USA,

[6] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801, USA,

*, †: These authors contributed equally.

§: To whom the correspondence should be addresses. Emails: zhao5@illinois.edu, milenkov@illinois.edu.

Abstract

Synthetic DNA-based data storage systems have received significant attention due to the promise of ultrahigh storage density and long-term stability. However, all proposed systems suffer from high cost, read-write latency and error-rates that render them impractical. One means to avoid synthesizing DNA is to use readily available native DNA. As native DNA content is fixed, one may adopt an alternative recording strategy that modifies the DNA topology to encode desired information. Here, we report the first macromolecular storage paradigm in which data is written in the form of "nicks (punches)" at predetermined positions on the sugar-phosphate backbone of native dsDNA. The platform accommodates parallel nicking on multiple "orthogonal" genomic DNA fragments, paired nicking and disassociation for creating "toehold" regions that enable single-bit random access and strand displacement computations. As a proof of concept, we used the programmable restriction enzyme *Pyrococcus furiosus* Argonaute to punch files into the PCR products of *Escherichia coli* genomic DNA. The encoded data is reliably reconstructed through simple read alignment.

Introduction

Existing DNA-based data recording architectures store user content in synthetic DNA oligos (1-12) and retrieve desired information via next-generation (e.g., HiSeq and MiSeq) or nanopore sequencing technologies (6). Although DNA sequencing can be performed routinely and at low cost, *de novo* synthesis of DNA strands with a predetermined content is a major bottleneck (15); DNA synthesis protocols add one nucleotide per cycle and are inherently slow and prohibitively expensive compared to existing optical and magnetic writing mechanisms. To address these limitations of DNA-based data storage systems and reduce their cost, we developed a new storage paradigm that represents information via *in vitro* topological modifications on native DNA sequences (e.g., genomic DNA or its cloned or PCR-amplified products).

In the write component of the proposed system (Figure 1, top), binary user information is converted into a positional code that describes where native DNA sequence is to be topologically modified, i.e. nicked. A nick is a cut in the sugar-phosphate backbone between two adjacent nucleotides in double-stranded DNA, and each nick encodes either $\log_2 2 = 1$ bit (if only one strand is allowed to be nicked or left unchanged) or $\log_2 3 = 1.58$ bits (if either of the two strands is allowed to be nicked or both left unchanged). As bacterial cells are easy to handle and grow, the native DNA nicking substrates of choice are the PCR products of one or multiple regions of bacterial genomic DNA, that can be easily isolated via simple and inexpensive available protocols. Native DNA is organized into *orthogonal registers*, with each register represented by multiple replicas of one isolated genomic region; two registers are termed orthogonal if their sequence edit distance is sufficiently large (>55%). Each register is nicked in a combinatorial fashion, determined by the information content to be stored. To enable fast and efficient data recording, a library of registers with desired nicking site patterns is created in parallel. Registers or orthogonal registers are subsequently placed into grids of microplates that enable random access to registers and spatially organize the data, similar to tracks and sectors on disks and tapes.

In the read component of the proposed system (Figure 1, bottom), nicked DNA is processed using next-generation sequencing (MiSeq) and the positions of nicks are determined via read analysis and subsequent reference-based sequence alignment.

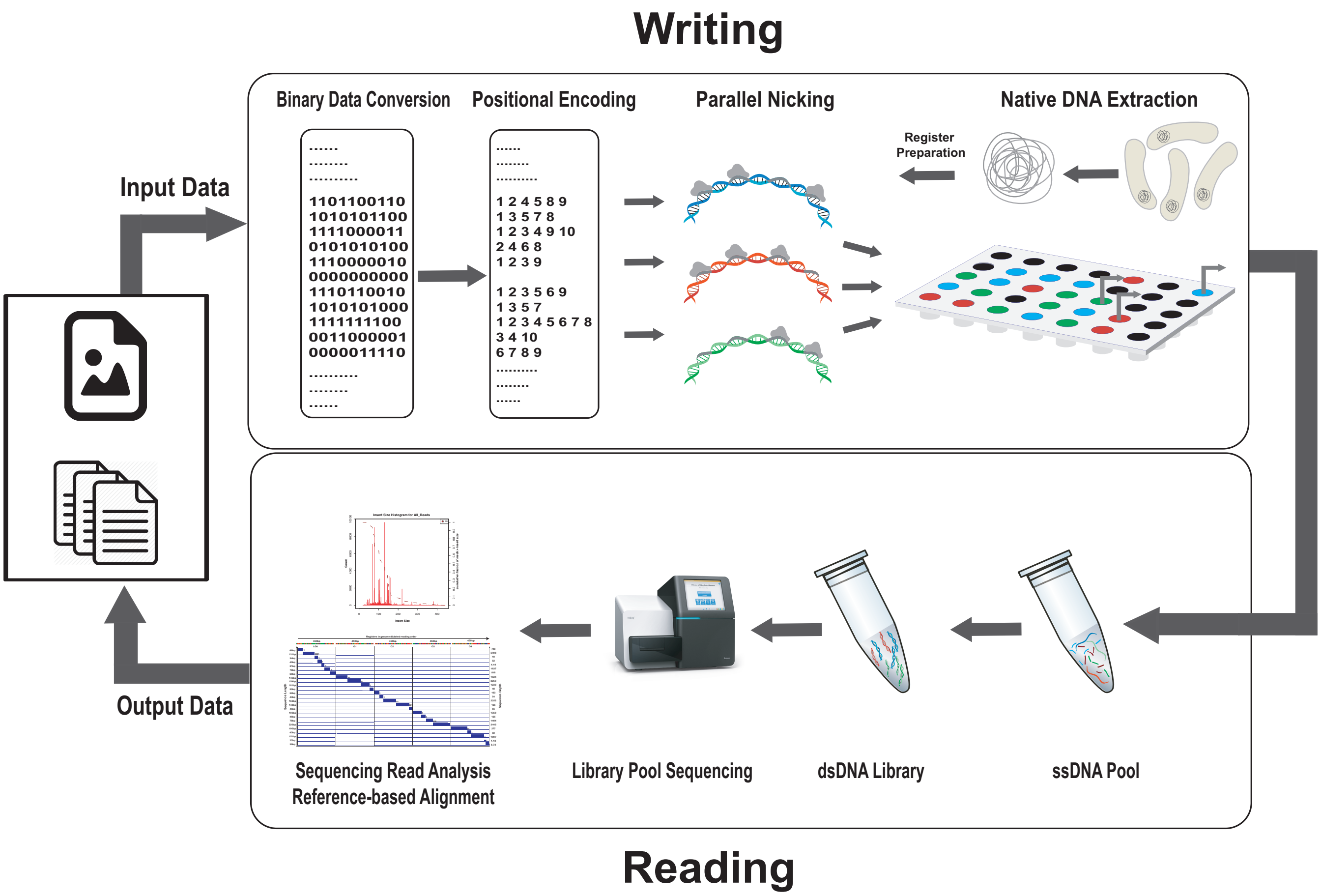

Figure 1 | The native DNA-based data storage platform. In the **Write component**, arbitrary user content is converted into a binary message. The message is then parsed into blocks of *m* bits, where *m* corresponds to the number of nicking positions on the register (for the running example, *m* = 10). Subsequently, binary information is translated into positional information indicating where to nick. Nicking reactions are performed in parallel via combinations of *Pf*Ago and guides. In the **Read component**, nicked products are purified and denatured to obtain a pool of ssDNAs of different lengths. The pool of ssDNAs is sequenced via MiSeq. The output reads are processed by first performing reference-based alignment of the reads, and then using read coverages to determine the nicked positions.

The register chosen for experimental verification is a DNA fragment of length 450 bps that was PCR-amplified from the genomic DNA of *E. coli* K12 MG1655. The register contains ten designated nicking positions. Although registers as long as 10 Kbps can be easily accommodated, they are harder to read; hence, multiple orthogonal registers are preferred to long registers. The nicking positions are determined based on four straightforward to accommodate sequence composition constraints (Supplementary Information; Section B.1) that enable precise nicking. To prevent disassociation of the two strands at room temperature, the nicking sites are placed at a conservative distance of at least 25 bps apart. The user file is parsed into 10-bit strings which are converted into nicking positions of spatially arranged registers, according to the rule that a '1' corresponds to a nick, while a '0' corresponds to the absence of a nick (the number of bits recorded is chosen based on the density of nicks and the length of the register). As an example, the string 0110000100 is converted into the positional code 238, indicating that nicking needs to be performed at the 2nd, 3rd and 8th positions (Figure 2a). Note that recording the bit '0' does not require any reactions, as it corresponds to the "no nick" ground state. Therefore, nick-based recording effectively reduces the size of the file to be actually recorded by half. This property of native DNA storage resembles that of compact disk (CD) and other recorders.

As the writing tool, we needed to choose a nicking enzyme with optimized programmability and nicking activity. Nicking endonucleases (natural/engineered) are only able to detect specific sequences in DNA strands; they can bind certain nucleotide sequences. Also, *Streptococcus pyogenes* Cas9 nickase (*Sp*Cas9n), as a widely used tool for genetic engineering applications, requires the presence of a protospacer adjacent motif (PAM) sequence (NGG) at the 3' site of the target DNA. The NGG motif constraint limits the nicking space to 1/16 of the available positions. The *Sp*Cas9n complex uses RNA guides (gRNAs) to bind the target, which makes it unstable and hard to handle. Furthermore, *Sp*Cas9n is a single turnover enzyme (16), i.e., one molecule of the enzyme can generate one nick per DNA molecule only. These make *Sp*Cas9n exhibit low efficiency and versatility for storage applications. To address these problems, we used the programmable restriction enzyme *Pyrococcus furiosus* Argonaute (*Pf*Ago) (13) as our writing tool. *Pf*Ago has significantly larger flexibility in double-stranded DNA cleaving than *Sp*Cas9n and, most importantly, has a high turnover rate (one enzyme molecule can be used to create a large number of nicks). *Pf*Ago also uses 16 nt DNA guides (gDNAs) that are more stable and easier to handle *in vitro*. We experimentally demonstrated that under proper reaction conditions, *Pf*Ago can successfully perform simultaneous nicking of multiple prescribed sites with high efficiency and precision within 40 min. A comparison of the nicking performance of *Sp*Cas9n and *Pf*Ago may be found in Table S2 and Figure S3-4.

To facilitate writing multiple user files in parallel, we designed *Pf*Ago guides for all ten nicking positions in the chosen register and created registers bearing all $2^{10} = 1024$ nicking combinations (Table S3). Registers were placed in microplates in an order dictated by the content to be encoded. The recording protocols for orthogonal registers, nick placements on both the sense and antisense strands and combinatorial mixing via group testing are described in the Supplementary Information (Figures S8, S10 and S17).

Since the length, sequence composition and nicking sites of a register are all known beforehand, reading amounts to detecting the positions of the nicks. The nicked registers are first denatured, resulting in ssDNA fragments of variable length dictated by the nicked positions. These length-modulated ssDNA fragments are subsequently converted into a dsDNA library, sequenced on Illumina MiSeq, and the resulting reads are aligned to the known reference register sequence. The positions of the nicks are determined based on read coverage analysis, the insert size distributions and through alignment with the reference sequence; potential nicking sites that are not covered are declared to be '0's (Figure 2a-c).

Results

As a proof of concept, we report write-read results for a 272-word file of size 0.4 KB containing Lincoln's Gettysburg Address (LGA) and a JPEG image of the Lincoln Memorial of size 14 KB (Figure S6). Both files were compressed and converted into ASCII and retrieved with perfect accuracy. Given the inherent redundancy of the sequencing process and the careful selection of the nicking sites and register sequences, no error-correction redundancy was needed (Figure 2b, c and Figure S5b-d). Technical details regarding implementations with orthogonal registers and with nicks on both DNA strands are provided in the Supplementary Information.

A faster, portable and more cost-effective method for reading the nicked DNA registers is via two-dimensional (2D) solid-state nanopore membranes. One approach is to use toeholds, short single-stranded regions on dsDNA created through two closely placed nicks, instead of single nicks. Experimental evidence reveals that toeholds can be accurately read using solid-state SiN_x and MoS_2 nanopores, as recently reported in (14). The cost of creating toeholds is twice as high as that of nicks, since one needs two different nicking guides. To mitigate this problem, one may attempt to detect nicks directly. To illustrate the feasibility of this approach, we performed Molecular Dynamics (MD) simulations based on quantum transport calculations. These revealed a strong inverse correlation between the ionic and electronic sheet current signals along the membrane induced by nicks in MoS_2 nanopores (Figures S12-S14 & Video S1). The simulation results reveal that solid-state nanopores may be able to detect information bearing nicks as well.

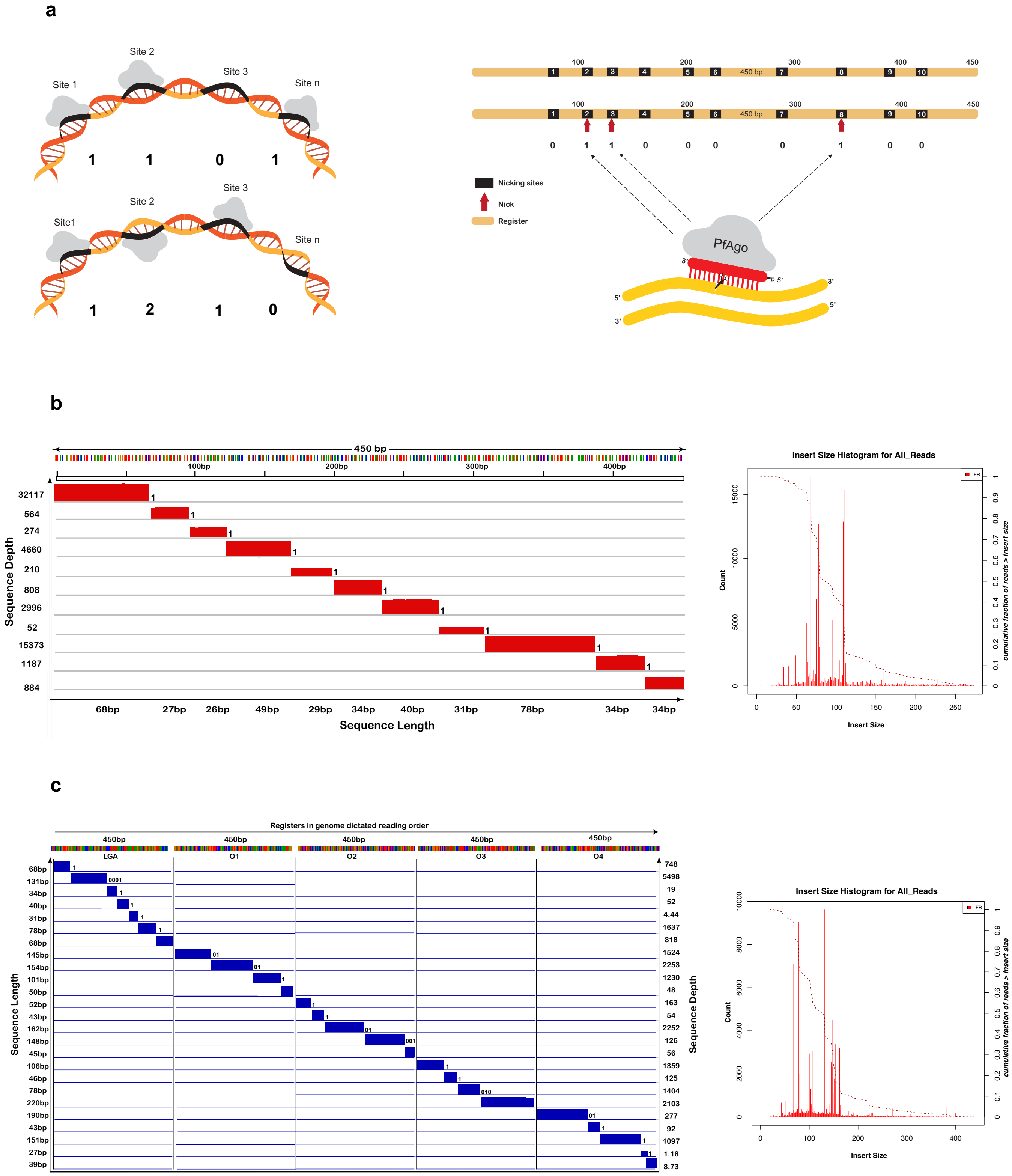

Figure 2 | Writing and reading the encoded data. **a**) *Pf*Ago can nick several pre-designated locations on only one strand (left, **top**) or both strands (left, **bottom**), simultaneously. In the first register, the stored content is 110…1, while in the second register, the content is 1210…0. The chosen register is a PCR product of a 450 bp *E. coli* genomic DNA fragment with 10 pre-designated non-uniformly spaced nicking positions. The positional code 238 corresponds to the binary vector 0110000100 (right). **b**) The MiSeq sequencing reads were aligned to the reference register to determine the positions of the nicks. The size distribution histogram (right) and coverage plots (left) are then generated based on the frequency and coverage depth of the reads. Coverage plots allow for straightforward detection of nicked and unnicked sites. In the example shown, all the ten positions were nicked, resulting in eleven aligned fragments. **C**) Five orthogonal registers used instead of one single register. Each vertical section represents one register in genome dictated reading order, and each row shows the read lengths retrieved after sequencing analysis. Read lengths are recorded on the left and sequencing depths on the right axis.

In addition to allowing for nanopore-based reading, toeholds also enable complex in-memory computations and for the first time, bitwise random access. In the former setting, toehold-mediated DNA strand displacement is a versatile tool for engineering dynamic molecular systems and performing molecular computations (17-19). Information is processed through releasing strands in a controlled fashion, with toeholds serving as initiation sites to which these input strands bind to displace a previously bound output strand. In the latter context, a toehold may represent a binary or non-binary symbol.

Toeholds are usually generated by binding two regions of synthetic ssDNA and leaving a short fragment unbound. However, with *Pf*Ago, one can easily create toeholds in native DNA. To form a toehold, two nicks are generated within 14 bps. Under appropriate buffer and temperature conditions, in a single reaction the 14 nt strand between the two nicks disassociates, leaving a toehold on the double-stranded DNA (Figure S15).

Fluorescence-based methods can detect the existence of a toehold (and hence enable symbol-wise random access) and estimate the concentration of registers bearing a toehold without modifying the DNA registers. We illustrate this process on a register encoding 0010000000, with a toehold of length 14 nts at the nicking position 3. As shown in Figure 3a, a fluorophore and quencher labelled reporter strand with a sequence complementary to the toehold can hybridize to the toehold segment, producing a fluorescence signal resulting from an increase of the distance between the fluorophore and the quencher. We were also able to reliably measure different ratios of native DNA fragments with and without toeholds within 20 mins (Figure 3b). Since the reporter has a short single stranded overhang, it can be pulled off from the register upon hybridization, making the readout process non-destructive (Polyacrylamide gel electrophoresis analysis, Figure 3c). This feature equips our proposed storage system with unique nondestructive bitwise random access, since one is able to design specific reporters to detect any desired toehold sequence which accompanies a nick. It also enables computations on data encoded in nicks, as described in two recent papers (20,21).

Discussion

In summary, by reprogramming *Pf*Ago as a universal nickase and using *E. coli* native DNA sequences, we have implemented the first DNA-based storage system that mitigates the use of costly long synthetic DNA strands for storing user information. Our platform utilizes a parallel writing mechanism that combines an inexpensive nicking enzyme and a small number of short and inexpensive synthetic DNA guides. In addition, this approach enables enzyme driven toehold creation, allowing for bitwise random access and in memory computing via strand displacement on data stored in the format of nicks.

Nick-based storage outperforms known synthetic DNA technologies in all relevant performance categories except for recording density; but the roughly one order of magnitude loss is insignificant for a system that already compacts petabytes in grams and overcompensated by the three to four-fold reduction of cost in our proposed system (Table 1; also, see Supplementary Information; Section B.10.). It also allows for cost-efficient scaling as a: long registers and mixtures of orthogonal registers may be nicked simultaneously; b: most uncompressed data files do not contain all possible 10-mers or compositions of orthogonal *k*-mers; c: genomic DNA and *Pf*Ago, as the writing tool, are readily available, and the mass of the created DNA products by far exceeds that of synthetic DNA, significantly increasing the number of readout cycles with NGS devices. This storage system may also be used to superimpose, erase and rewrite categorical and metadata on synthetic DNA oligos, in which case bitwise random access enables efficient non-destructive search and concentration sensing.

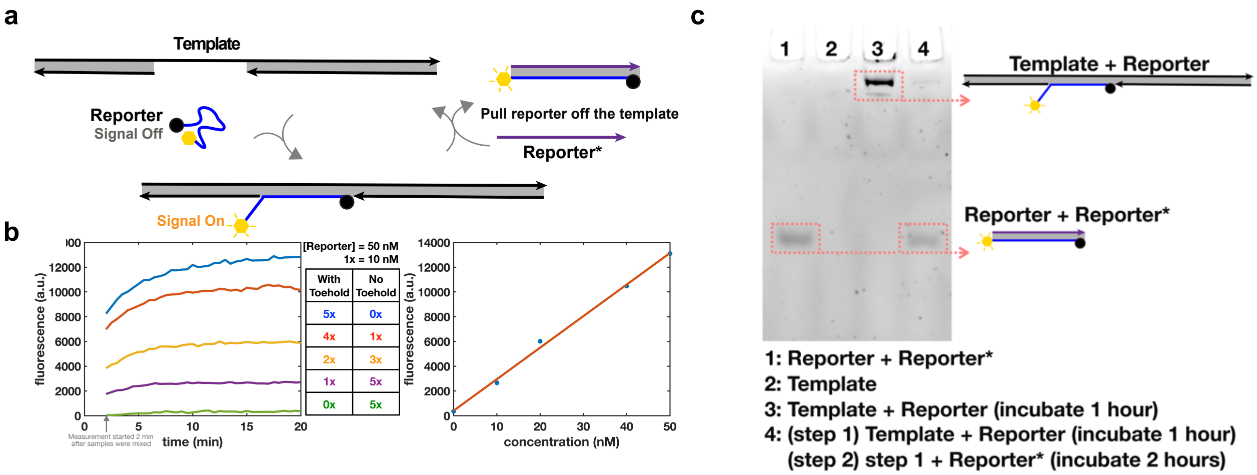

Figure 3 | Non-destructive bitwise random access. **a)** Non-destructive detection of toeholds through a fluorophore and quencher labelled Reporter strand. Once the Reporter hybridizes with the toehold on the register strand, a fluorescence signal is observed due to the increase of the distance between the fluorophore and quencher. The Reporter strand can be pulled off from the register once the Reporter* strand hybridizes with the Reporter. **b)** Kinetics of detecting the concentrations of registers with and without toeholds in a mixtures **(**left**).** The fluorescence signals saturate within 20 minutes. The samples were mixed no more than 2 min before measurement. The concentration of toehold-ed DNA can be accurately quantified through fluorescence intensity **(**right**),** as it increases linearly with the concentration of the registers with toehold. **c)** PAGE gel results for non-destructive detection of a toehold. The gel was not stained with other fluorescence dyes, thus only the species with self-fluorescence is observed. After adding the Reporter, a large size complex appears in lane 3, indicating hybridization of the Reporter and the register. After the Reporter* is added, as seen in lane 4, the large size complex in lane 3 no longer exhibits self-fluorescence, indicating that the Reporter strand is pulled off from the register.

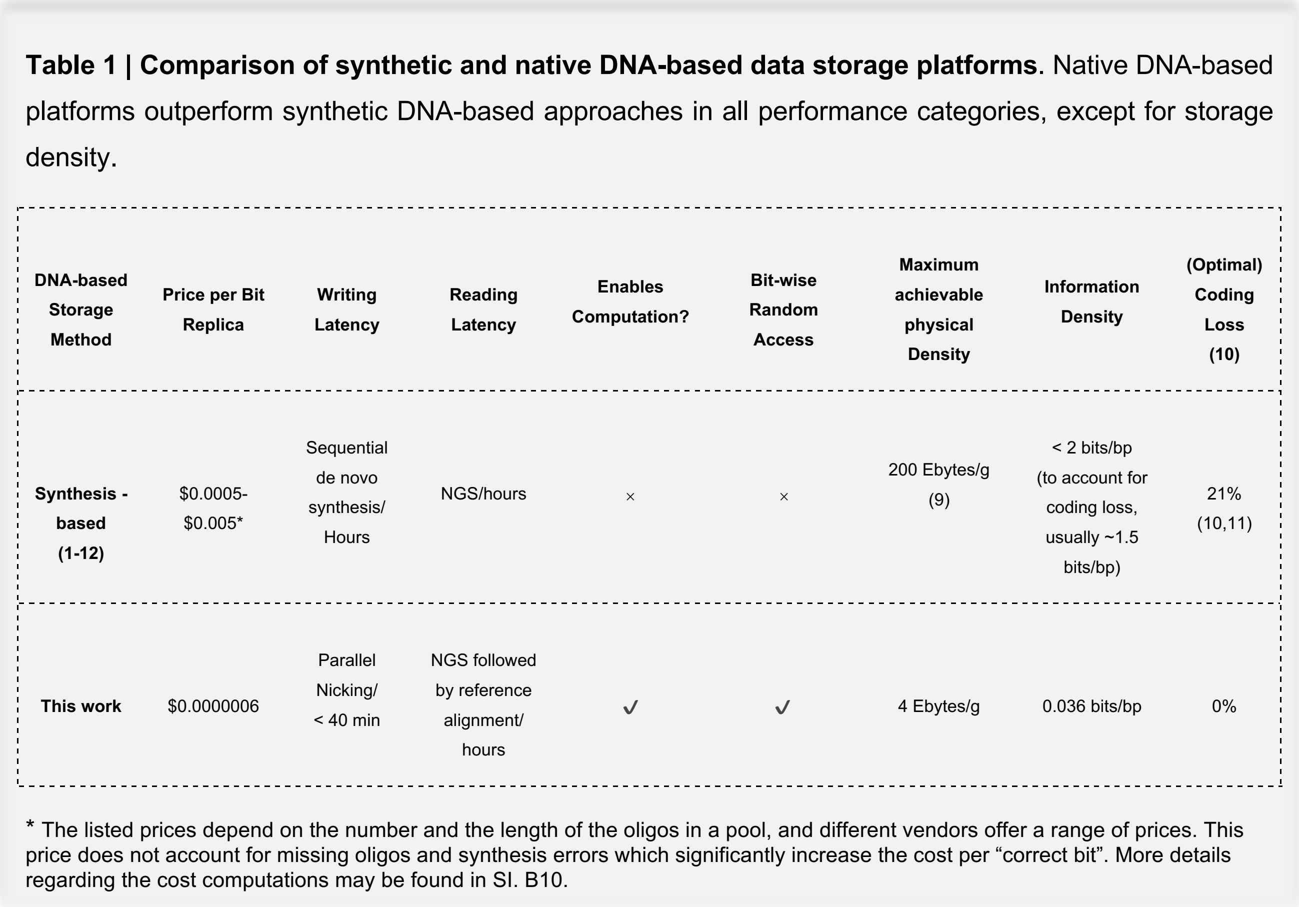

Table 1 | Comparison of synthetic and native DNA-based data storage platforms. Native DNA-based platforms outperform synthetic DNA-based approaches in all performance categories, except for storage density.

DNA-based Storage Method	Price per Bit Replica	Writing Latency	Reading Latency	Enables Computation?	Bit-wise Random Access	Maximum achievable physical Density	Information Density	(Optimal) Coding Loss (10)
Synthesis - based (1-12)	$0.0005-$0.005*	Sequential de novo synthesis/ Hours	NGS/hours	×	×	200 Ebytes/g (9)	< 2 bits/bp (to account for coding loss, usually ~1.5 bits/bp)	21% (10,11)
This work	$0.0000006	Parallel Nicking/ < 40 min	NGS followed by reference alignment/ hours	✓	✓	4 Ebytes/g	0.036 bits/bp	0%

* The listed prices depend on the number and the length of the oligos in a pool, and different vendors offer a range of prices. This price does not account for missing oligos and synthesis errors which significantly increase the cost per “correct bit”. More details regarding the cost computations may be found in SI. B10.

Materials and Methods

Guide DNA selection and positional coding

To minimize off-site nicking rates, increase the efficiency and accuracy of gDNA-binding, and eliminate readout errors, the nicking regions were selected by imposing the following constraints: each region is of length 16 bps, which allows each gDNA to bind to a unique position in the chosen registers; each region has a GC content in the range 20-60%; there are no G repeats longer than three (this feature is only needed in conjunction with nanopore sequencing); the gDNAs are at Hamming distance at least eight from each other; and the nicking sites are placed at least 25 bps apart. Positional coding is performed based on the number of orthogonal registers used, and the number of nicking positions selected on each register. For the single-register implementation with one-sided nicking, ten positions were selected on a 450 bp genomic fragment, bearing 10 bits.

Although choosing longer registers with a larger number of nicking positions is possible and indeed easily doable, we selected the given length for our proofs of concept in order to accommodate different sequencing technologies. The five orthogonal register implementation may encode 32 bits with one sided nicking, and roughly 50 bits with two-sided nicking. Hence, each binary encoded message is parsed into blocks of length either 10, or 32 or 50 bits, which are recorded via nicking.

Genomic DNA isolation and PCR amplification

Genomic DNA was extracted from an overnight culture of *E. coli* K12 MG1655, using the Wizard® Genomic DNA Purification Kit (Promega). The kit can be used for at least 100 isolations. One extraction yields up to 100 µg of genomic DNA (from 5 ml overnight culture) which can be used for several hundreds of amplification reactions. Isolated genomic DNA was subsequently stored at 4 °C. DNA amplification was performed via PCR using the Q5 DNA polymerase and 5X Q5 buffer (New England Biolabs) in 50 µl. All primers purchased from Integrated DNA Technologies (IDT). In all PCR reactions, 10-50 ng of *E. coli* genomic DNA and 25 pmol of forward and reverse primers were used. The PCR protocol consists of: 1) 3 min at 98 °C, 2) 20 s at 98 °C, 3) 20 s at 62 °C, 4) 15 s at 72 °C, 5) go to step 2 and repeat the cycle 36 times, 6) 8 min at 72 °C. Each PCR reaction produced ~2-2.5 µg of the register string, sufficient for >100 reactions. PCR products were run on 1% agarose gel and purified using the Zymoclean gel DNA recovery kit (Zymo Research).

Enzyme expression and purification

Enzyme expression and purification was performed as previously described (13). More than 200 nmols of *Pf*Ago were purified from 1 L of *E. coli* culture, enabling >50,000 reactions.

*Pf*Ago nicking experiments

For ease of access and spatial organization of data, pools of registers are kept in 384-well plates. The distribution of the registers and enzymatic reagents was performed manually (when transferring volumes of reagents with volumes of the order of microliters) or using the Echo® 550 liquid handler (LABCYTE) (when transferring minute volumes of reagents of the order of nanoliters). The latter allows for testing the reaction efficiency of *nanoliters* of reagents and it also represents a faster and more efficient manner of liquid handling at larger scales. *Pf*Ago reactions were performed in buffer conditions including 2 mM $MnCl_2$ 150 mM NaCl, and 20 mM HEPES, pH 7.5, and a total volume of 10-50 μL. After adding the buffer, dsDNA registers, ssDNA phosphorylated gDNAs and the enzyme, the sample was thoroughly mixed by pipetting 6-8 times. Nicking was performed based on the following protocol: 1) 15 min at 70 °C, 2) 10 min at 95 °C, 3) gradual decrease of temperature (0.1 °C/s) to 4 °C. In all reactions, we used 3.75-5 pmol of *Pf*Ago and 20-50 ng of the register. gDNAs were either phosphorylated using T4 Polynucleotide Kinase (NEB) in lab or phosphorylated guides were purchased from IDT. For each nicking reaction, a (2-10):1 ratio of guides to enzymes was formed. All guides were used in equimolar mixtures.

Cas9 nickase experiments

The Cas9 D10A nickase was purchased from IDT (Alt-R® S.p. Cas9 D10A Nickase); crRNAs were designed via IDT's Custom Alt-R® CRISPR-Cas9 guide RNA design tool. Both crRNAs and tracrRNAs were purchased from IDT and hybridized based on the manufacturer's protocol. The 10x Cas9 reaction buffer included: 200 mM HEPES, 1M NaCl, 50 mM $MgCl_2$, 1mM EDTA, pH 6.5. All Cas9n nicking reactions were set-up based on the manufacturer's protocol and performed at 37 °C for 60 min.

Protocol verification via gel electrophoresis

ssDNA gel analysis was performed using a 2% agarose gel. Nicked dsDNA samples were first denatured at high temperature (99 °C) for 10 min, and immediately cooled to 4 °C. The ssDNA products were then run on a pre-made 2% agarose Ex-Gel (Thermo Fisher Scientific).

Sample preparation for MiSeq sequencing

All nicked PCR products (obtained either via *Pf*Ago or Cas9n reactions) were purified using the Qiaquick PCR purification kit (QIAGEN) and eluted in ddH_2O. The dsDNA registers were denatured at 99 °C for 10 min, and immediately cooled down to 4 °C. The ssDNA samples were first quantified via the Qubit 3.0 fluorometer. Next, the Accel-NGS® 1S plus DNA library kit (Swift Biosciences) was used for library preparation following the manufacturer's recommended protocol. Prepared libraries were quantitated with Qubit, and then run on a DNA Fragment Analyzer (Agilent, CA) to determine fragment sizes, pooled in equimolar concentration. The pool was further quantitated by qPCR. All steps were performed for each sample separately and no nicked DNA samples were mixed.

MiSeq sequencing

The pooled libraries were loaded on a MiSeq device and sequenced for 250 cycles from each end of the library fragments with a Nano V2 500 cycles kit (Illumina). The raw fastq files were generated and demultiplexed with the bcl2fastq v2.20 Conversion Software (Illumina).

Reference alignment

Data was processed using a Nextflow-based workflow (22), implemented as follows. Sequence data was trimmed using Trimmomatic v.0.36 (23) in paired-end mode using the options "ILLUMINACLIP: adapters/TruSeq3-PE-2.fa:2:15:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:20". Reads were aligned to the reference sequence using bwa v 0.7.10 (24) with the command "bwa mem -t 12 <REFERENCE> <R1> <R2>". Alignments were sorted and processed using samtools v1.6 (25). Insert size statistics were collected using Picard v.2.10.1 (26). Aligned files (BAMs) were then split based on expected fragment size using sambamba (27) with the option "sambamba view -t 4 -f bam -h -F "(template_length >= [LOWER] and template_length <= [UPPER]) or (template_length >= -[UPPER] and template_length <= -[LOWER])", with the upper and lower bound settings in brackets originally set to allow for one additional base greater and lesser than the expected size. Read coverage files were then generated using bedtools (28) and bedGraphToBigWig (29). Alignment and coverage information was visualized in IGV v2.3.10 (30). All the scripts used for data analysis are available from the corresponding authors upon request.

Nanopore simulations

To obtain the trajectories of the nicked molecule translocating through the nanopore, all-atom Molecular Dynamics simulations were performed using NAMD (31). For these simulations, the DNA structure (30 nucleotides around the 5th nicking site in the register) was obtained from the 3D-DART webserver (32) and described using the CHARMM27 force field (33). Appropriate backbone molecules were manually removed to create the nicks in the desired locations of the strand. Note that in order to obtain a stronger nanopore current signal from the DNA backbone, the PO_3 groups located at the nicked position may be removed by treatment of the nicked dsDNA with a phosphatase enzyme such as BAP (bone alkaline phosphatase).

The DNA molecule was placed just above the nanopore of a Molybdenum disulfide (MoS_2) membrane to ensure a successful translocation process. The nanopore membrane and the biomolecule were then solvated in a water box with ions (K^+ and Cl^-) placed randomly to reach a neutrally charged system of concentration 1 M. Van der Waals energies were calculated using a 12 Å cutoff. Each system was minimized for 5000 steps and further equilibrated for 2 ps in an NPT ensemble, where the system was maintained at 1 atm pressure by a Langevin Piston (34) and at constant 300 K temperature using a Langevin thermostat. After equilibration, an external electric field was applied to the system in vertical direction to drive the nicked DNA through the nanopores.

A trajectory file of molecules driven through the nanopore by the applied electric field obtained from the MD simulations was used to calculate the ionic current via Equation (1) (35), where q_i and z_i denote the charge and z-coordinate of ion *i*, respectively; *V* denotes the voltage bias (1 V) and *L* the length of the water box along the z-direction, while *N* represents the number of ions and *Δt* the interval between the trajectory frames:

$$I(t)=\frac{1}{\Delta t.L_z}\sum_{i=1}^{N}q_i\left(z_i\left(t+\Delta t\right)-z_i\left(t\right)\right)$$

For each frame of the trajectory, the electrostatic potential is calculated using the following non-linear Poisson Boltzmann formula

$$\nabla.[\varepsilon(r)\nabla\varphi(r)]=-e[C_{K^+}(r)-C_{Cl^-}(r)]-\rho_{DNA}(r),$$

where ρ_{DNA} denotes the charge density of DNA, $\varepsilon(r)$ the local permittivity, and where $C_{K^+}(r)$ and $C_{Cl^-}(r)$ equal the local electrolyte concentrations of K^+ and Cl^- and obey the Poisson-Boltzmann statistics. The detailed description of the method used is outlined elsewhere (36). The calculated electrostatic potential is used to obtain the transverse sheet conductance in MoS_2 quantum point contact nanopore membranes. The electronic transport is formulated as a self-consistent model based on the semi-classical thermionic Poisson-Boltzmann technique using a two-valley model within the effective mass approximation. The calculated conductance at a given energy mode is described according to

$$G_{n1,2}=\frac{2e^2}{h}\frac{1}{1+\exp\left(\frac{E^K_{n1,2}-E^L_F}{K_BT}\right)}+\frac{2e^2}{h}\frac{1}{1+\exp\left(\frac{E^Q_{n1,2}-E^L_F}{K_BT}\right)}$$

where E_F^L denotes the quasi-Fermi level and is set depending on the carrier concentration (chosen to be 10^{12} cm^{-2}); in addition, $n_{1,2}$ represents the energy modes of the two conductance channels while $E^K_{n1,2}$ and $E^Q_{n1,2}$ stand for the energy modes at these two channels caused by the effective masses K and Q, respectively. A detailed discussion of the thermionic current model is described elsewhere (37).

As a final remark, we observe that the simulations in Figure S12-S14 indicating strong negative correlations of the global minimum and maximum of the sheet and ion current may be interpreted as follow: When nicked DNA translocates through the pore, the oscillations of the nicked backbone allow more ions to pass through the pore, leading to a steep increase (maximum) in the ion current. At the same time, the absence of the PO_3 group charges leads to a decrease in the sheet current to its global minimum.

References and Notes

1. Skinner. G.M, Visscher. K, Mansuripur. M, Biocompatible writing of data into DNA, *Journal of Bionanoscience*, **1**, 1-5 (2007).

2. Church. G. M., Gao. Y, Kosuri. S. Next-generation digital information storage in DNA. *Science* **337,** 1628-1628 (2012).

3. Goldman. N, Bertone. P, Chen. S, Dessimoz. C, Leproust. E, Sipos. B, Birney. E. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494,** 77-80 (2013).

4. Yazdi. S. H. T, Yuan. Y, Ma. J, Zhao. H, Milenkovic. O. A rewritable, random-access DNA-based storage system. *Sci. Rep.* **5,** 14138 (2015).

5. Grass. R. N., Heckel. R., Puddu. M., Paunescu. D. & Stark. W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew. Chem. Int. Ed.* **54,** 2552–2555 (2015).

6. Yazdi. S. H. T., Gabrys. R., and Milenkovic. O. "Portable and error-free DNA-based data storage." *Scientific reports* 7.1 (2017).

7. Shipman. S. L., Nivala. J., Macklis. J. D. & Church. G. M. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* **547,** 345–349 (2017).

8. Zhirnov. V., Zadegan. R. M., Sandhu. G. S., Church. G. M. & Hughes. W. L. Nucleic acid memory. *Nat. Mater.* **15,** 366-370 (2016).

9. Erlich. Y. & Zielinski. D. DNA fountain enables a robust and efficient storage architecture, *Science*, **355**, 950-954 (2017).

10. Yazdi. S. H. T, Kiah. HM, Ruiz-Garcia.E, Ma. J, Zhao. H, Milenkovic. O. DNA-based storage: Trends and Methods. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications,* **1, 3,** 230-248, (2015).

11. Laure. C, Karamessini. D, Milenkovic, O, Charles. L, Lutz. J.F, Coding in 2D: using intentional dispersity to enhance the information capacity of sequence-coded polymer barcodes. *Angewandte Chemie International Edition*, ***55***(36), pp.10722-10725 (2016).

12. Milenkovic. O., Gabrys. R., Kiah. H. M, Yazdi. S. H. T., Exabytes in a Test Tube. *IEEE Spectrum*, *55*(5), 40-45 (2018).

13. Enghiad. B., Zhao. H. Programmable DNA-guided artificial restriction enzymes. ACS Synth. Biol., **6**, 752−757 (2017).

14. Liu. K., Pan. C, Kuhn. A, Nievergelt. A.P, Fantner. G, Milenkovic. O, Radenovic. A, Detecting topological variations of DNA at single-molecule level, *Nature Communications*, **10**, 3 (2019).

15. Palluk. S, Arlow. D. H, de Rond. T, Barthel. S, Kang. J. S, Bector. R, Baghdassarian. H. M, Truong. A. N, Kim. P.W, Singh. A. K, Hillson. N. J, Keasling. J. D., De novo DNA synthesis using polymerase- nucleotide conjugates*, Nat Biotechnol.*, **36**, 645-650 (2018).

16. Andres. C, Jinek. M. In vitro enzymology of Cas9, *Methods Enzymol.* **546**, 1-20 (2016).

17. Yurke. B, Turberfield. A.J, Mills. A.P, Simmel. F.C, Neumann. J.L, A DNA-fueled molecular machine made of DNA. *Nature*, **406**:605–608 (2000).

18. Zhang. DY., Seelig. G, Dynamic DNA nanotechnology using strand-displacement reactions. *Nat Chem.,* **3**:103–113 (2011).

19. Wang. B., Thachuk. C, Ellington. A., Winfree. E., Soloveichik. D., Effective design principles for leakless strand displacement systems, *PNAS*, **115** (52), E12182-E12191 (2018).

20. Wang. B., Chalk. C, Soloveichik. D, SIMDNA: Single Instruction, Multiple Data Computation with DNA Strand Displacement Cascades *DNA 25 Conference,* Seattle*,* WA, U.S.A. (2019).

21. Chen. T, Riedel. M, Parallel Binary Sorting and Shifting with DNA, *11th International Workshop on Bio-Design Automation (IWBDA),* Cambridge, England, U.K. (2019).

22. Di Tommaso, P., et al., Nextflow enables reproducible computational workflows*. Nat Biotechnol*, **35**(4) 316-319 (2017).

23. Bolger, A.M., M. Lohse, and B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**(15): p. 2114-20 (2014).

24. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv e-prints,1303 (2013).

25. Li, H., et al., The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16): p. 2078-9 (2009).

26. Institute, B. *Picard Tools*. [2018 2017]; Available from: http://broadinstitute.github.io/picard/.

27. Tarasov, A., et al., Sambamba: fast processing of NGS alignment formats. Bioinformatics, **31**(12): p. 2032-4 (2015).

28. Quinlan, A.R., BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics*, **47**: p. 11 12 1-34, (2014)

29. Kent, W.J., et al., BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**(17): p. 2204-7 (2010).

30. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, **14**(2): p. 178-92 (2013).

31. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26,** 1781–1802 (2005).

32. Van Dijk, M. & Bonvin, A. M. J. J. 3D-DART: a DNA structure modelling server. *Nucleic Acids Res.* **37,** W235–W239 (2009).

33. Foloppe, N. & MacKerell, Jr., A. D. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.* **21,** 86–104 (2000).

34. Feller, S. E., Zhang, Y., Pastor, R. W. & Brooks, B. R. Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.* **103,** 4613–4621 (1995).

35. Aksimentiev, A., Heng, J. B., Timp, G. & Schulten, K. Microscopic Kinetics of DNA Translocation through Synthetic Nanopores. *Biophys. J.* **87,** 2086–2097 (2004).

36. Girdhar, A., Sathe, C., Schulten, K. & Leburton, J.-P. Graphene quantum point contact transistor for DNA sensing. *Proc. Natl. Acad. Sci.* **110,** 16748–16753 (2013).

37. Sarathy, A. & Leburton, J. P. Electronic conductance model in constricted MoS_2 with nanopores. *Appl. Phys. Lett.* **108** (2016).

Supplementary Materials:

Figures S1-S17

Tables S1-S5

Video S1

References

Acknowledgements This work was funded by the DARPA W911NF-18-2-0032 Molecular Informatics program. We thank Christopher Fields for performing the MiSeq read analysis and reference alignments, and SMH Tabatabaei Yazdi, and Jianhao Peng for many discussions and help.

Author contributions O.M., H.Z. and SK.T. developed the nicking-based data recording platform. SK.T. performed the nicking and toehold creation experiments. D.S., B.W., O.M. and SK.T. designed the bitwise random-access system. D.S. and B.W. performed the bitwise access experiments. JP.L. and N.A. designed the nanopore simulations. A.G.H., D.S., O.M. and SK.T. designed the readout system, while A.G.H. performed MiSeq sequencing. B.E. performed initial *Pf*Ago nicking activity verifications and helped with nicking experiment designs and protein purifications.

Competing interests O.M., H.Z., A.G.H. and SK.T. have filed a patent on native DNA-based data storage via nicking**.**

Correspondence and requests for materials should be addressed to H.Z or O.M.