ACS SyntheticBiology

pubs.acs.org/synthbio Letter

Reading and Writing Digital Information in TNA

Kefan Yang, Cailen M. McCloskey, and John C. Chaput*

Cite This: ACS Synth. Biol. 2020, 9, 2936–2942 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: DNA has become a popular soft material for low energy, high-density information storage, but it is susceptible to damage through oxidation, pH, temperature, and nucleases in the environment. Here, we describe a new molecular chemotype for data archiving based on the unnatural genetic framework of α-L-threofuranosyl nucleic acid (TNA). Using a simple genetic coding strategy, 23 kilobytes of digital information were stored in DNA-primed TNA oligonucleotides and recovered with perfect accuracy after exposure to biological nucleases that destroyed equivalent DNA messages. We suggest that these results extend the capacity for nucleic acids to function as a soft material for low energy, high-density information storage by providing a safeguard against information loss caused by nuclease digestion.

KEYWORDS: *TNA, XNA, information storage, nuclease resistance*

As data storage struggles to keep pace with data generation, molecular approaches provide an attractive alternative to traditional magnetic-based devices for long-term rare-access storage.[1] DNA in particular has garnered significant interest as a natural material capable of low energy, high-density information storage for data archiving.[2] At a theoretical maximum of 2 bits per nucleotide (nt), one gram of DNA can store 455 exabytes of information,[3] which is equivalent to 25% of the annual worldwide data production.[4] As the cost of DNA synthesis and sequencing continues to fall,[5,6] the capacity for digital information storage has steadily risen from a modest 35 bit image produced in 1996 to the recent demonstration of a massive ∼200 MB assembly of 35 files.[7,8] DNA processing has also achieved major successes with new methodologies for error correction,[9] maximization of information density,[10] and selective access to files.[8] Progress in the area of biomaterials has led to a DNA-of-things (DoT) storage system for producing objects that carry information, including a blueprint for their own synthesis.[11] Finally, the commercial market for DNA coding and information systems is growing with DNA-tagging systems available from companies like SelectaDNA, Selectamark, and DNA Technologies.

Despite its many advantages, DNA is susceptible to information loss through accidental or unexpected encounters with DNA degrading enzymes (nucleases) present in the environment. Whereas inorganic matrices have been explored as a strategy for thermal stability,[12] approaches for safeguarding against nuclease digestion are critically lacking. Here we describe the use of α-L-threofuranosyl nucleic acid (TNA) as an artificial genetic system for next-generation information storage. TNA is an unnatural genetic polymer (Figure 1a) in which the natural five-carbon deoxyribose sugar found in DNA has been replaced with an unnatural four-carbon threose sugar.[13] This chemical change leads to a 2′,3′-linked phosphodiester backbone structure that is recalcitrant to nuclease digestion.[14] With the discovery of engineered polymerases that are able to convert genetic information back and forth between DNA and TNA,[15,16] we hypothesized that TNA offered a biologically stable alternative to DNA as a soft material for data archiving if TNA could be shown to replicate molecular messages with high fidelity.

"The quick brown fox jumps over 13 lazy dogs." is a common phrase used to evaluate new information storage systems because it contains each letter in the English alphabet along with a numerical value. To test the capacity for TNA to function as a soft material for information storage, we converted this phrase from 47 bytes of digital information in binary UTF-8 format into 16 unique 83-nt strands of DNA using the following equivalencies: A, C = 0 and G, T = 1 (Supplementary Figure S1 and Supplementary Tables S1, S2).[3] Each DNA strand consists of a central 24-bit data block, a 19-bit address specifying the location of the data block, and conserved flanking regions for PCR amplification and sequencing (Supplementary Figure S2). A complete cycle of write−store−read (Figure 1b) was performed by transcribing the DNA strands *en masse* into TNA, purifying the TNA by denaturing polyacrylamide gel electrophoresis (PAGE), and reverse transcribing the TNA back into DNA for PCR amplification and sequencing. As TNA is an artificial genetic

Received: July 7, 2020
Published: September 23, 2020

https://dx.doi.org/10.1021/acssynbio.0c00361
ACS Synth. Biol. 2020, 9, 2936–2942

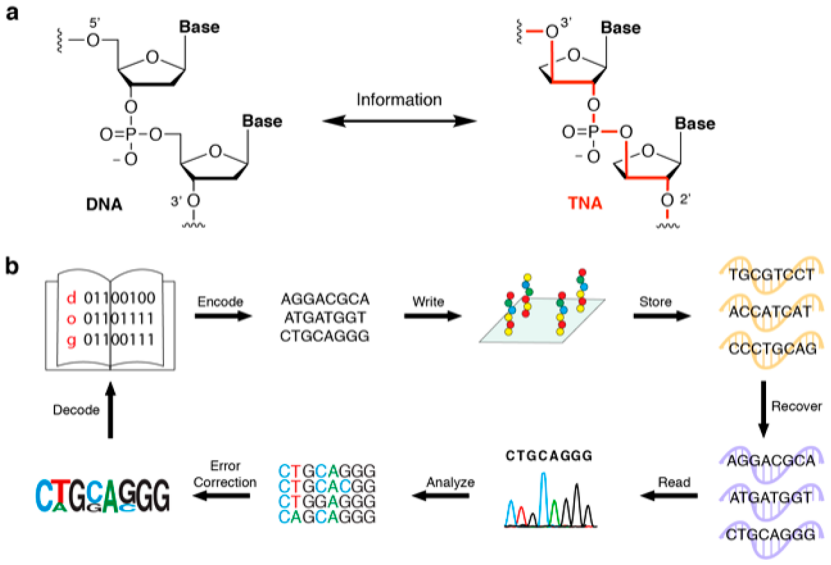

Figure 1. Overview of TNA data storage. (a) Constitutional structures for the linearized backbone of DNA (left) and α-L-threofuranosyl-(3′,2′) nucleic acid, TNA (right). TNA has a five-atom backbone repeat unit with quasi *trans*-diaxial positioning of the 3′,2′ phosphodiester linkages. (b) Reading and writing digital information in TNA. Each write−store−read cycle involves encoding digital information in DNA, writing the information in DNA format, converting the information into TNA, recovering the information in DNA format, reading the information by NGS sequencing, analyzing the data, and decoding the message.

polymer, the writing process requires enzymatic synthesis using chemically synthesized TNA triphosphates (tNTPs)[17] and a laboratory-evolved TNA polymerase[18] that copies DNA templates into TNA; however, TNA recovery needs only DNA triphosphates (dNTPs) and a natural DNA polymerase from a thermophilic bacteria (*Geobacillus stearothermophilus*).[19,20]

Sanger sequencing of 56 clones recovered from the write−store−read cycle yielded complete recovery of the test phrase. We used an "AA" watermark that only appears in TNA strands that are reverse transcribed back into DNA to confirm that the sequenced DNA originated from TNA templates (Supplementary Figure S3). Sequence analysis shows that the process occurs with 99.2% fidelity with G → C transversion accounting for the majority (5.3 per 1000 nt incorporation events) of the genetic changes (Supplementary Figure S4). This error mode occurs during TNA synthesis when tGTP nucleotides adopt a *syn* conformation with respect to the glycosidic bond to form a Hoogsteen base pair with dG residues in the template (Supplementary Figure S5).[21] However, this problem can be overcome by preparing the tNTP mixture with the unnatural nucleotide analogue, 7-deaza-tGTP, which lacks the nitrogen heteroatom at the *N*-7 position necessary to form a tGTP:dG Hoogsteen base pair in the polymerase active site. Repeating the write−store−read cycle with the new tNTP mixture increased the fidelity to 99.8% (Supplementary Figure S4).

Encouraged by the success of the test phrase, we demonstrated the capacity for high density information storage by writing 22 349 bytes of digital information in TNA. We chose the Declaration of Independence and the logo for the University of California, Irvine (UCI) as representative formats for text and picture files, respectively. We converted the data into a bit stream that was encoded in 7451 unique DNA oligonucleotides. The DNA library was synthesized by array-based inkjet printing on DNA microchips, amplified by PCR, and made single-stranded. The information was faithfully written into TNA and stored as single-stranded oligonucleotides following PAGE purification. To read the encoded information, the TNA was reverse transcribed back into cDNA, PCR amplified, and sequenced using an Illumina MiSeq platform. We decoded the information by joining the overlapping paired-end reads, removing reads that were either shorter than the expected 83-nt length or lacked a perfect barcode sequence, and aligned the merged reads with the reference index. For each data block, we generated a consensus sequence with >500-fold average coverage (Figure 2). All of the data blocks were recovered with enough redundancy that each nucleotide had an error rate of <0.5%. By simulation, we estimate that 100% of the information could be recovered from 10% of the total reads (Figure 2). To our knowledge, this was the most comprehensive (largest number of templates) fidelity study performed on any XNA system developed to date.

We performed a parallel study with DNA to compare the efficiency of TNA storage to known DNA systems. This study evaluated the quality of the reads obtained from a write−store−read cycle performed on the same 23 kB of information. Interestingly, the results were quantitatively similar for DNA and TNA with 83.9% and 78.6% of the reads producing the correct length sequence, respectively (Figure 3a). Deletions were the major error mode for both classes of nucleic acids (14.1% and 20.1% for DNA and TNA, respectively), which is likely due to the limitations of array-based DNA synthesis. At the base level, the fidelity of the reads was also congruent, yielding values of 99.4% and 99.2% for DNA and TNA, respectively (Figure 3b). Likewise, the pattern of substitutions, insertions, and deletions was similar between DNA and TNA. The distribution of errors with regard to strand position is also similar for both systems, with most errors occurring at the first nucleotide addition step adjacent to the conserved primer binding sites (Figure 3c). The increase in errors at the first position was attributed to strand switching between the DNA primer and elongated TNA product, which has been observed previously with other synthetic genetic polymers.[16] Moreover, nearest neighbor analysis indicate that very little sequence bias

https://dx.doi.org/10.1021/acssynbio.0c00361
ACS Synth. Biol. 2020, 9, 2936−2942

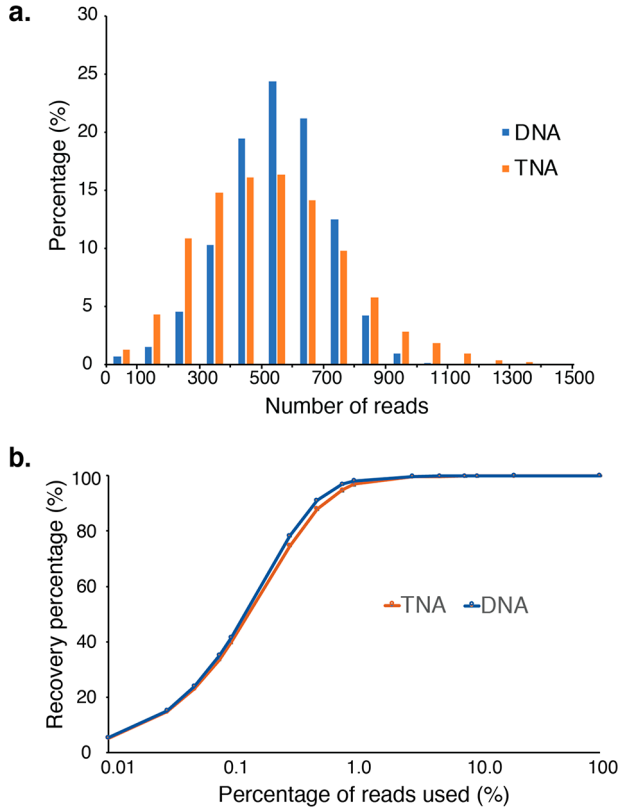

Figure 2. Coverage analysis and recovery simulation. (a) Read distribution observed for the high-density DNA (blue) and TNA (orange) encoded libraries. The average coverage for DNA and TNA is 551X and 533X, respectively. (b) Simulation of information recovery shows that all information could be fully recovered from 10% of the total reads from either the DNA or TNA information systems.

exists between the DNA and TNA information storage systems when 7-deaza-tGTP is present in the nucleotide mixture used for transcription of the TNA strands (Figure 4).

To establish TNA as a biologically stable system for information storage, we incubated the 23 kB data blocks encoding the Declaration of Independence and UCI logo files with human serum, human liver microsomes, and snake venom phosphodiesterase. After 24 h of incubation at 37 °C, sample analysis by denaturing PAGE reveals that TNA remains undigested while the DNA samples were completely destroyed (Figure 5a). Importantly, the information encoded in TNA samples exposed to biological nucleases was recovered in its entirety despite its synthesis from a DNA primer (Figure 5b). While the TNA portion of the oligonucleotide protects the DNA primer from nucleolytic attack by 3′ exonucleases, we suspect that base pairing within and between TNA strands may account for enhanced stability of the DNA primer against other types of nucleases. However, greater stability, if needed, can always be achieved by performing the writing step with a TNA primer. In addition, we found that chimeric DNA−TNA heteroduplexes formed by copying DNA templates into TNA are also protected from nuclease digestion (Figure 5c). These striking observations highlight the value of alternative genetic polymers, like TNA, as informational macromolecules for nucleic acid memory, either as materials for information storage or as molecular sentinels for protecting DNA information.

Recognizing that increased storage capacity requires the use of longer TNA strands, we performed a proof-of-principle demonstration of TNA synthesis on longer DNA templates. After 3 h of incubation at 55 °C, sample analysis by denaturing PAGE reveals that TNA strands up to 200 nt in length can be produced by Kod-RSGA (Supplementary Figure S6). Although the yield of the 200 nt TNA product is ∼20-fold less than that of the 83 nt TNA product used in the current study, the 500-fold coverage provided by the 83 nt system suggests that the use of longer TNA strands is a viable option without loss of information. However, if needed, the process of storing information in TNA polymers can be further improved through polymerase engineering.[18]

Although artificial genetic polymers capable of heredity, replication, and evolution are still in their infancy,[16] advances in this area have grown at a steady pace.[18,22] General strategies established to synthesize nucleoside triphosphates on the gram scale will help provide researchers with ample supplies of unnatural building blocks.[17] Similar progress in ultrahigh throughput microfluidic-based screening has made it possible to create custom engineered polymerases that function with high efficiency and fidelity.[15] Because these new molecular biology tools rely on DNA as a basis for writing and reading molecular information, they will continue to benefit from new advances made in the synthesis and sequencing of DNA. Through further progress, one could imagine applying TNA to embedded memory (DoT) systems to produce medical implants or wearable sensors that report on a persons' health status or exposure to toxins.[11] Steganography offers another possible application, in which messages are hidden in everyday objects. Relative to DNA, XNAs provide an additional layer of security, as the engineered polymerase must be available and correctly matched with its class of nucleic acid polymer. This is especially true for new examples of XNA polymerases that have not been described in the literature [*e.g.*, xylo nucleic acid (XyloNA) or glycerol nucleic acid (GNA)], which could be considered as an analogous methodology to asymmetric cryptography. Here, the encoding process (DNA into XNA) would offer a private key, as the specific polymerase variant and XNA triphosphates are not commercially available, while the subsequent decoding process would offer a public key, as DNA polymerases and dNTP substrates are widely available.

In summary, our study extends the emerging field of DNA-based information storage to include artificial genetic polymers by demonstrating the methodology for reading and writing digital information in TNA. We suggest that these findings help derisk the use of nucleic acid polymers as molecular storage systems by safeguarding against information loss through nuclease digestion. We suggest that TNA provides an exciting opportunity for exploring unnatural polymers as molecules for nucleic acid memory. Our strategy for information storage could be applied generally to other classes of nucleic acid polymers with physicochemical properties that are distinct from those found in natural DNA and RNA.

METHODS

General Information. SurePrint oligonucleotide libraries were purchased from Agilent Technologies (Santa Clara, CA). Nuclease-free water was purchased from Invitrogen (Thermo Fisher Scientific, US). Single DNA oligonucleotides were purchased from Integrated DNA Technologies (IDT, Coralville, IA), purified by denaturing urea polyacrylamide gel electrophoresis (PAGE), electroeluted, buffer exchanged, and

https://dx.doi.org/10.1021/acssynbio.0c00361
ACS Synth. Biol. 2020, 9, 2936−2942

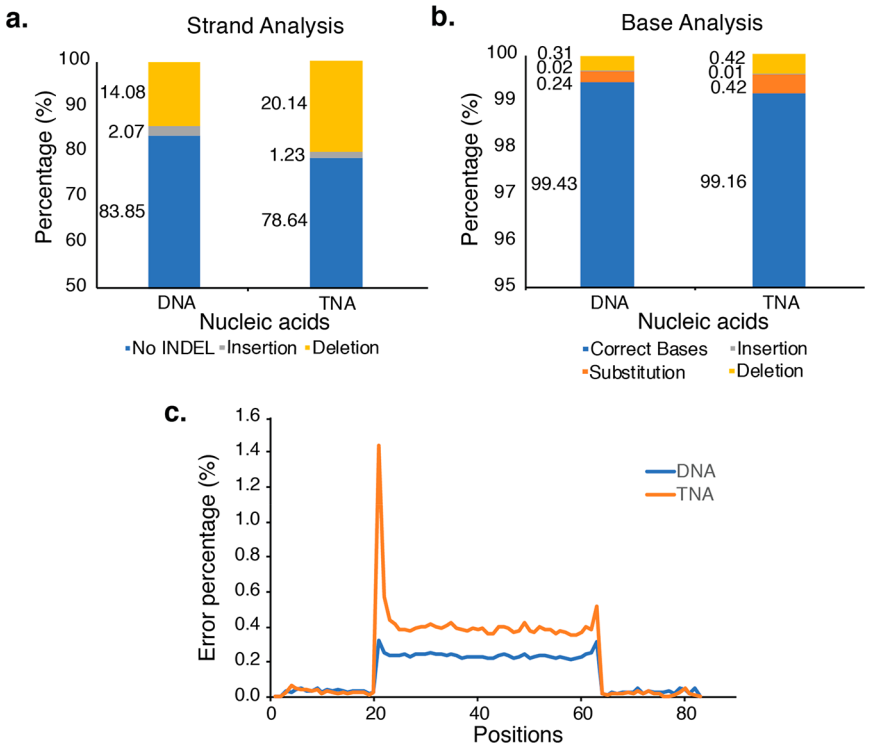

Figure 3. Nucleotide-based error analysis and position-based error analysis. (a) Strand analysis. 83.9% and 78.6% of the DNA and TNA reads, respectively, are of the correct length. Deletions were the major error mode (14.1% and 20.1% for DNA and TNA, respectively). Insertions occurred less often in TNA than DNA, (1.2% *vs* 2.1%). (b) Base composition analysis. Among all reads, 99.4% and 99.2% of base-reads are correct for DNA and TNA, respectively. (c) The distribution of errors with regard to strand position is similar for both systems, with most errors occurring at the first and last nucleotide addition steps during primer extension.

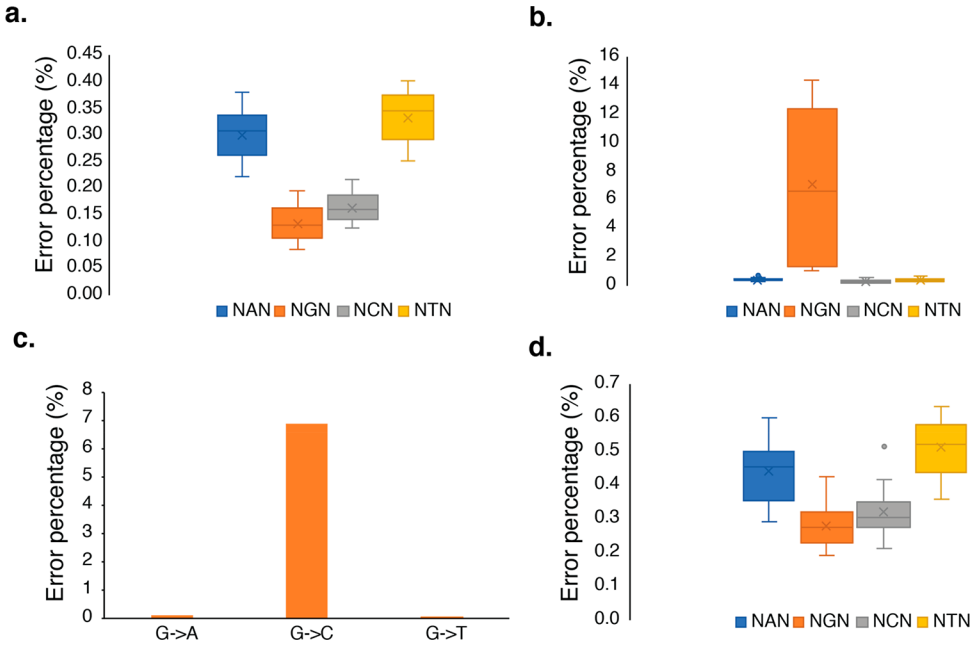

Figure 4. Nearest neighbor analysis. (a) Error analysis of the original DNA-encoded information from the DNA control group. (b) Error analysis of the TNA information system using the standard tNTP mixture (tATP, tCTP, tGTP, tTTP) with natural bases. (c) Specific error mode analysis observed for G nucleotides in the TNA information storage using the standard tNTP mixture. (d) Error analysis of the TNA information system using the modified tNTP mixture (tATP, tCTP, tTTP, and 7-deaza-tGTP) to eliminate the error mode caused by tGTP-dG mispairing. Abbreviations: N, degeneracy of all 4 nucleotides (A, T, C, G).

concentrated using Millipore YM-10 or YM-30 Centricon centrifugal filter units, and quantified by UV absorbance *via* NanoDrop (Thermo Fisher Scientific, Waltham, MA). Taq DNA polymerase, Proteinase K, and 10× ThermoPol buffer were purchased from New England Biolabs (NEB, Ipswich, MA). Kod-RSGA TNA polymerase was expressed and purified

https://dx.doi.org/10.1021/acssynbio.0c00361
ACS Synth. Biol. 2020, 9, 2936−2942

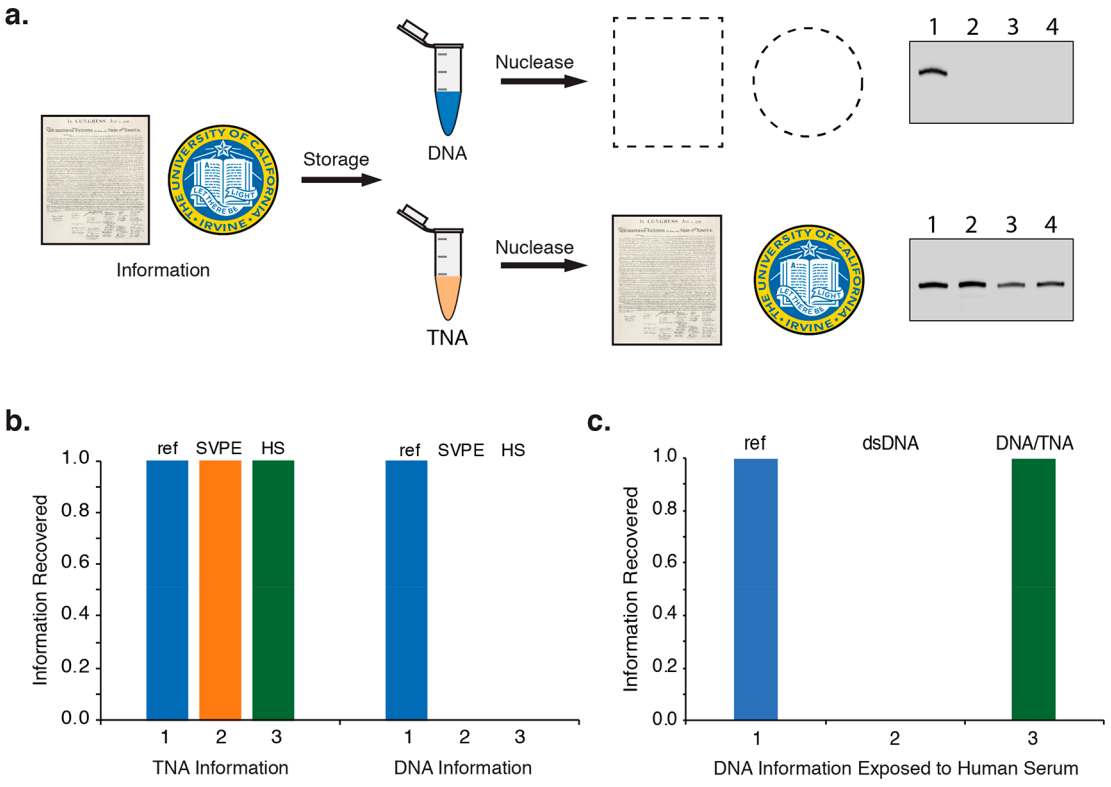

Figure 5. File recovery from nuclease exposure. (a) DNA and TNA oligonucleotides encoding 23 kB information were analyzed by denaturing PAGE after exposure to biological nucleases for 24 h at 37 °C. Nuclease-free control (lane 1), SVPE (lane 2), HS (lane 3), HLM (lane 4). (b) TNA information recovered after nuclease exposure, reverse transcription, PCR amplification, and DNA sequencing. (c) DNA information recovered after nuclease exposure, PCR amplification, and DNA sequencing. Abbreviations: snake venom phosphodiesterase (SVPE), human serum (HS), human liver microsomes (HLM). DNA/TNA refers to a chimeric duplex in which one strand is DNA and the complementary strand is TNA. The UCI seal was provided courtesy of the University of California, Irvine for use in this figure.

as described previously.[23] DNA triphosphates were purchased from Sigma-Aldrich (St. Louis, MO). TNA triphosphates bearing natural bases and 7-deaza tGTP were synthesized as previously described.[17,21] Sequences were cloned into TOPO vectors using TOPO-TA Cloning Kit, with PCR 2.1 Vector (Thermo Fisher Scientific) and transformed into DH5α competent cells (NEB).

Data Representation and Encoding. The test phrase "The quick brown fox jumps over 13 lazy dogs." was written in a .txt file and converted to binary using UTF-8 encoding. The binary was encoded into DNA using A or C for 1 and T or G for 0, as described previously.[3] Each strand of DNA is 83 nt in length with 19 nt for the index and 24 nt for the data payload. The payload consists of 3 bytes of information using an 8 nt/byte conversion from binary to DNA. The remaining 40 nt are reserved for the forward (PBS7) and reverse (PBS8) primer binding sites (Table S1). The sequences were purchased as 16 separate DNA strands. We used the logo for the University of California—Irvine and The Declaration of Independence as examples of the types of information that can be stored in TNA. The UCI logo image was downloaded from https://brand.uci.edu/brand-assets/marks/. The file size was reduced to 10 217 bytes using the html base64 format and then converted to binary. The Declaration of Independence (US, 1776) was downloaded from https://www.ushistory.org/declaration/document/, and 12 132 bytes were stored in .txt format and converted to binary. All binary files were encoded in DNA using the conversion described above for the test phrase. Together, both files totaling 22 349 bytes were encoded in 7451 DNA strands that were synthesized by Agilent Technologies using array-based DNA synthesis.

Information Storage Using the Test Phrase. The 16 DNA strands encoding the test phrase were resuspended in nuclease-free water. Aliquots were mixed in equal volume to a final concentration of 20 μM. The DNA templates were copied into TNA using an engineered TNA polymerase in a primer extension reaction. The primer extension reaction (300 μL volume) contained 1 μM of FAM-labeled PBS9-PBS8 DNA primer, 1 μM of mixed DNA templates, 1× ThermoPol buffer, 0.1 mM tNTPs, and 2 μM Kod-RSGA TNA polymerase.[17,18] The primer was annealed to the DNA templates and the reaction was initiated by adding the tNTPs and polymerase. After incubation at 55 °C for 3 h, the samples were washed with phenol/chloroform and dried under reduced pressure. TNA strands were purified by 10% denaturing urea PAGE. Gels were visualized using a Typhoon biomolecular imager. Bands corresponding to full-length product were cut and electroeluted at 75 V overnight and desalted using a YM-30 centrifugal filter before drying under reduced pressure. To recover the information stored, the single-stranded in TNA was reverse transcribed using Bst DNA polymerase and the resulting cDNA was sequenced. The reaction was performed in a volume of 12 μL containing at least 1 μM TNA template, 1× ThermoPol buffer, 400 μM dNTPs, 35 μM primer PBS7, 5 mM $MgCl_2$, 2 μM of Bst DNA polymerase (LF). Reactions were annealed and initiated by adding dNTPs and Bst DNA polymerase. After incubating at 50 °C for 3 h, the polymerase was digested with 1 μL 800 units/mL of Proteinase K for 30

https://dx.doi.org/10.1021/acssynbio.0c00361
ACS Synth. Biol. 2020, 9, 2936−2942

min at 50 °C, which was heat inactivated for 30 min at 95 °C. cDNA was then amplified by PCR in a 50 μL volume containing 1× ThermoPol, 400 μM dNTPs, 1 mM $MgSO_4$, 1 μM of each primer PBS9 and PBS7 using 1.25 units/50 μL Taq polymerase. After TOPO cloning and colony PCR, colony plasmids were sent out for Sanger sequencing. The results were aligned with reference templates.

Information Storage Using the UCI Logo and the Declaration of Independence. The DNA oligonucleotide pool was resuspended in 50 μL nuclease-free water. The library was PCR amplified using the PEGylated PBS7 primer and PBS9-PBS8 primer, which allow for strand separation by denaturing PAGE and addition of the primer binding site for TNA synthesis. The PCR reaction was carried out in a 5 mL volume containing 1× ThermoPol, 400 μM dNTPs, 1 mM $MgSO_4$, 1.25 units/50 μL Taq polymerase, and 0.1 fmol/μL 200-fold dilution of the resuspended library. The reaction was dried completely and denatured with 95% formamide with 25 mM EDTA for 10 min at 95 °C before loading on a 10% denaturing urea PAGE gel. Bands corresponding to the template strands were cut and electroeluted at 75 V overnight. For the DNA control, the oligo pool was PCR amplified using the PBS7 primer and PBS8 primer in 50 mL volume containing 1× ThermoPol, 400 μM dNTPs, 1 mM $MgSO_4$, 1.25 units/50 μL Taq polymerase, and 0.1 fmol/μL 200-fold dilution of the resuspended library.

TNA Synthesis. TNA strands were synthesized by primer extension. The primer extension (300 μL) contained 1 μM IR680-labeled PBS9, 1 μM amplified DNA template, 1× ThermoPol buffer, 0.1 mM tATP, tTTP, tCTP, and 7-deaza tGTP, and 2 μM Kod-RSGA. Samples are annealed and the reaction was initiated by adding the tNTPs and polymerase. After incubation at 55 °C for 3 h, the samples were washed with phenol/chloroform and dried under reduced pressure. TNA strands were purified by 10% denaturing urea PAGE. Gels were visualized using a LI-COR Odyssey CLx imager. The correct bands were cut and electroeluted at 75 V overnight and desalted using a YM-30 centrifugal filter before drying under reduced pressure.

DNA Synthesis for Nuclease Digestion. The DNA sample used for nuclease digestion was prepared by primer-extension (300 μL) using 1 μM IR680-labeled PBS9, 1 μM amplified DNA template, 1× ThermoPol buffer, 0.1 mM dNTPs for each, and 1 μM Kod-wild type polymerase. Samples are annealed and the reaction was initiated by adding the dNTPs and polymerase. After incubation at 55 °C for 3 h, the samples were washed with phenol/chloroform and dried under reduced pressure. Control DNA strands were purified by 10% denaturing urea PAGE. Gels were visualized using a LI-COR Odyssey CLx imager. The correct bands were cut and electroeluted at 75 V overnight and desalted using a YM-30 centrifugal filter before drying under reduced pressure.

Reverse Transcription of TNA Back into cDNA. To recover the information stored, the single-stranded in TNA was reverse transcribed using Bst DNA polymerase and the resulting cDNA was sequenced. The reaction was performed in a volume of 12 μL containing at least 1 μM TNA template, 1× ThermoPol buffer, 400 μM dNTPs, 35 μM primer PBS7, 5 mM $MgCl_2$, 2 μM of Bst DNA polymerase (LF). Reactions were annealed and initiated by adding dNTPs and Bst DNA polymerase. After incubating at 50 °C for 3 h, the polymerase was digested with 1 μL 800 units/mL of Proteinase K for 30 min at 50 °C, which was heat inactivated for 30 min at 95 °C. cDNA was then amplified by PCR in a 50 μL volume containing 1× ThermoPol, 400 μM dNTPs, 1 mM $MgSO_4$, 1 μM of each primer PBS9 and PBS7 using Taq polymerase.

TOPO Cloning and Colony PCR. The amplified cDNA was transformed into DH5α Competent Cells through TOPO-TA cloning. After growing cells on LB-ampicillin plates overnight, colonies were picked off the plate, PCR amplified, and verified on 2% agarose gels.

Next Generation Sequencing Analysis. cDNA (200 ng) from the TNA sample and 200 ng of DNA amplicon from the DNA control sample were sequenced with Illumina's MiSeq platform using 75 bp paired-end protocol. The raw sequencing data was first trimmed using Trimmomatic v.0.35[24] and then the overlapping reads were combined to a single contig using SeqPrep (https://github.com/jstjohn/SeqPrep). The contigs were aligned to the reference index using Bowtie 2.[25] After alignment, data were analyzed through python for error rate, coverage and recovery rate. The raw data and codes are publicly available in Github: https://github.com/kefany1/TNA-information-storage/.

Biological Stability Challenge. Single-stranded TNA and DNA samples were subjected to a nuclease challenge. Ten pmol of TNA or control DNA were incubated in the presence of phosphodiesterase I (1 mU/μL) from *Crotalus adamanteus* venom (SVPE), 50% human serum (HS) or 2 mg/mL human liver microsomes (HLM) separately. All samples were incubated at 37 °C for 24 h. Afterward, the samples (1 pmol) were analyzed by 10% denaturing PAGE. Gels were visualized using a LI-COR Odyssey CLx imager.

Simulation. Using a standard python script, the number of unique reads was determined after random data samples (10.00%, 8.00%, 5.00%, 3.00%, 1.00%, 0.80%, 0.50%, 0.30%, 0.10%, 0.08%, 0.05%, 0.03%, 0.01%) were taken from the data set. For example, among all 7451 unique reads, 0.1% of data set gives 2931 unique reads, which yields a recovery rate of 39.3%.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.0c00361.

Tables S1−S2, Figures S1−S6 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

John C. Chaput − *Departments of Pharmaceutical Sciences, Chemistry, and Molecular Biology and Biochemistry, University of California, Irvine, California 92697-3958, United States;* orcid.org/0000-0003-1393-135X; Phone: 949-824-8149; Email: jchaput@uci.edu

Authors

Kefan Yang − *Departments of Pharmaceutical Sciences, Chemistry, and Molecular Biology and Biochemistry, University of California, Irvine, California 92697-3958, United States*

Cailen M. McCloskey − *Departments of Pharmaceutical Sciences, Chemistry, and Molecular Biology and Biochemistry, University of California, Irvine, California 92697-3958, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.0c00361

https://dx.doi.org/10.1021/acssynbio.0c00361
ACS Synth. Biol. 2020, 9, 2936−2942

Author Contributions

K.Y. and J.C. conceived of the project and designed the experiments. K.Y. and C.M. performed the experiments. J.C. and K.Y. wrote the manuscript. All authors reviewed and commented on the manuscript.

Notes

The authors declare no competing financial interest.

Codes and raw data are available in Github. https://github.com/kefany1/TNA-information-storage/.

■ ACKNOWLEDGMENTS

This work was supported by a grant from the National Science Foundation (MCB: 1946312) to J.C. We gratefully acknowledge the Genomic High Throughput Sequencing Facility, Jenney Wu, and Whitney England at UCI.

■ REFERENCES

(1) Cafferty, B. J., Ten, A. S., Fink, M. J., Morey, S., Preston, D. J., Mrksich, M., and Whitesides, G. M. (2019) Storage of Information Using Small Organic Molecules. *ACS Cent. Sci. 5*, 911−916.

(2) Ceze, L., Nivala, J., and Strauss, K. (2019) Molecular digital data storage using DNA. *Nat. Rev. Genet. 20*, 456−466.

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

(4) Shrivastava, S., and Badlani, R. (2014) Data storage in DNA. *Int. J. Electr. Engergy 2*, 119−124.

(5) Kosuri, S., and Church, G. M. (2014) Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods 11*, 499−507.

(6) Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet. 17*, 333−51.

(7) Davis, J. (1996) Microvenus + Biology, contemporary art, genetics. *Art J. 55*, 70−74.

(8) Organick, L., Ang, S. D., Chen, Y. J., Lopez, R., Yekhanin, S., Makarychev, K., Racz, M. Z., Kamath, G., Gopalan, P., Nguyen, B., Takahashi, C. N., Newman, S., Parker, H. Y., Rashtchian, C., Stewart, K., Gupta, G., Carlson, R., Mulligan, J., Carmean, D., Seelig, G., Ceze, L., and Strauss, K. (2018) Random access in large-scale DNA data storage. *Nat. Biotechnol. 36*, 242−248.

(9) Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., and Birney, E. (2013) Towards practical, high-capacity, low-maintenance information storagein syntheic DNA. *Nature 494*, 77−79.

(10) Erlich, Y., and Zielinski, D. (2017) DNA Fountain enables a robust and efficient storage architecture. *Science 355*, 950−954.

(11) Koch, J., Gantenbein, S., Masania, K., Stark, W. J., Erlich, Y., and Grass, R. N. (2020) A DNA-of-things storage architecture to create materials with embedded memory. *Nat. Biotechnol. 38*, 39−43.

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

(13) Schöning, K. U., Scholz, P., Guntha, S., Wu, X., Krishnamurthy, R., and Eschenmoser, A. (2000) Chemical etiology of nucleic acid structure: the a-threofuranosyl-(3′−>2′) oligonucleotide system. *Science 290*, 1347−1351.

(14) Culbertson, M. C., Temburnikar, K. W., Sau, S. P., Liao, J.-Y., Bala, S., and Chaput, J. C. (2016) Evaluating TNA stability under simulated physiological conditions. *Bioorg. Med. Chem. Lett. 26*, 2418−2421.

(15) Nikoomanzar, A., Vallejo, D., and Chaput, J. C. (2019) Elucidating the Determinants of Polymerase Specificity by Microfluidic-Based Deep Mutational Scanning. *ACS Synth. Biol. 8*, 1421−1429.

(16) Pinheiro, V. B., Taylor, A. I., Cozens, C., Abramov, M., Renders, M., Zhang, S., Chaput, J. C., Wengel, J., Peak-Chew, S. Y., McLaughlin, S. H., Herdewijn, P., and Holliger, P. (2012) Synthetic genetic polymers capable of heredity and evolution. *Science 336*, 341−344.

(17) Liao, J.-Y., Bala, S., Ngor, A. K., Yik, E. J., and Chaput, J. C. (2019) P(V) Reagents for the Scalable Synthesis of Natural and Modified Nucleoside Triphosphates. *J. Am. Chem. Soc. 141*, 13286−13289.

(18) Nikoomanzar, A., Chim, N., Yik, E. J., and Chaput, J. C. (2020) Engineering polymerases for applications in synthetic biology. *Q. Rev. Biophys. 53*, No. e8.

(19) Dunn, M. R., and Chaput, J. C. (2016) Reverse transcription of threose nucleic acid by a naturally occurring DNA polymerase. *ChemBioChem 17*, 1804−1808.

(20) Jackson, L. N., Chim, N., Shi, C., and Chaput, J. C. (2019) Crystal structures of a natural DNA polymerase that functions as an XNA reverse transcriptase. *Nucleic Acids Res. 47*, 6973−6983.

(21) Mei, H., Liao, J.-Y., Jimenez, R. M., Wang, Y., Bala, S., McCloskey, C., Switzer, C., and Chaput, J. C. (2018) Synthesis and Evolution of a Threose Nucleic Acid Aptamer Bearing 7-Deaza-7-Substituted Guanosine Residues. *J. Am. Chem. Soc. 140*, 5706−5713.

(22) Chaput, J. C., Yu, H., and Zhang, S. (2012) The emerging world of synthetic genetics. *Chem. Biol. 19*, 1360−1371.

(23) Nikoomanzar, A., Dunn, M. R., and Chaput, J. C. (2017) Engineered polymerases with altered substrate specificity: expression and purification. *Curr. Protoc. Nucleic Acid Chem. 69*, 4.75.

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

(25) Langmead, B., and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods 9*, 357−9.

https://dx.doi.org/10.1021/acssynbio.0c00361
ACS Synth. Biol. 2020, 9, 2936−2942