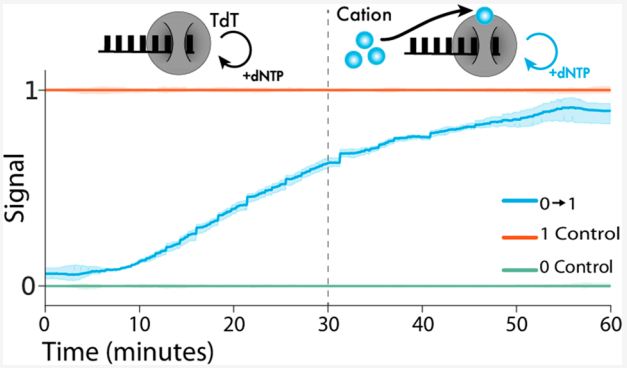
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

Recording Temporal Signals with Minutes Resolution Using Enzymatic DNA Synthesis

Namita Bhan,[□] Alec Callisto,[□] Jonathan Strutz, Joshua Glaser, Reza Kalhor, Edward S. Boyden, George Church, Konrad Kording, and Keith E. J. Tyo*

Cite This: J. Am. Chem. Soc. 2021, 143, 16630−16640
Read Online

ABSTRACT: Employing DNA as a high-density data storage medium has paved the way for next-generation digital storage and biosensing technologies. However, the multipart architecture of current DNA-based recording techniques renders them inherently slow and incapable of recording fluctuating signals with subhour frequencies. To address this limitation, we developed a simplified system employing a single enzyme, terminal deoxynucleotidyl transferase (TdT), to transduce environmental signals into DNA. TdT adds nucleotides to the 3′-ends of single-stranded DNA (ssDNA) in a template-independent manner, selecting bases according to inherent preferences and environmental conditions. By characterizing TdT nucleotide selectivity under different conditions, we show that TdT can encode various physiologically relevant signals such as Co^{2+}, Ca^{2+}, and Zn^{2+} concentrations and temperature changes *in vitro*. Further, by considering the average rate of nucleotide incorporation, we show that the resulting ssDNA functions as a molecular ticker tape. With this method we accurately encode a temporal record of fluctuations in Co^{2+} concentration to within 1 min over a 60 min period. Finally, we engineer TdT to allosterically turn off in the presence of a physiologically relevant concentration of calcium. We use this engineered TdT in concert with a reference TdT to develop a two-polymerase system capable of recording a single-step change in the Ca^{2+} signal to within 1 min over a 60 min period. This work expands the repertoire of DNA-based recording techniques by developing a novel DNA synthesis-based system that can record temporal environmental signals into DNA with a resolution of minutes.



INTRODUCTION

DNA is an attractive medium for both long-term data storage and for *in vitro* recording of molecular events due to its high information density[1−3] and long-term stability.[4] Molecular recording strategies write information into DNA by altering existing DNA sequences[5] or adding new sequences.[6] For example, systems have been developed that use methods including differential CRISPR spacer acquisition,[5,7,8] enzymatic synthesis,[9−11] and others.[1,8,12] By connecting these DNA modifications to a user input (in the case of data storage) or environmental signal of interest (in the case of recording events), these strategies enable *post hoc* recovery of signal dynamics over time by DNA sequencing. To date, molecular recording systems, both *in vitro* and *in vivo*, have connected signals of interest to DNA recordings with transcriptional control, using signal-responsive promoters to drive the expression of molecular writers, such as base-editors, CRISPR-associated systems, and gene-circuits, to record changes in signal. These approaches have yielded accurate recordings; however, the time required to transduce signals through a recording apparatus that includes transcription, translation, and DNA modification fundamentally constrains the application of these methods to events on the time scales of hours or days. A recording mechanism that relies only on post-translational elements would be inherently faster, as signal transduction would only require subsecond conformational shifts in one enzyme.

In an effort to speed up DNA recording processes, we hypothesized that a DNA polymerase (DNAp), which continually incorporates bases,[13] could serve as a candidate for post-translational molecular encoding. In such a system, a DNAp functions as a "ticker-tape" recorder, transforming changes in environmental signals into changes in the composition of the DNA it synthesizes[14] (Figure 1A). Much faster than transcription and translation, nucleotide incorpo-

Received: July 14, 2021
Published: September 30, 2021

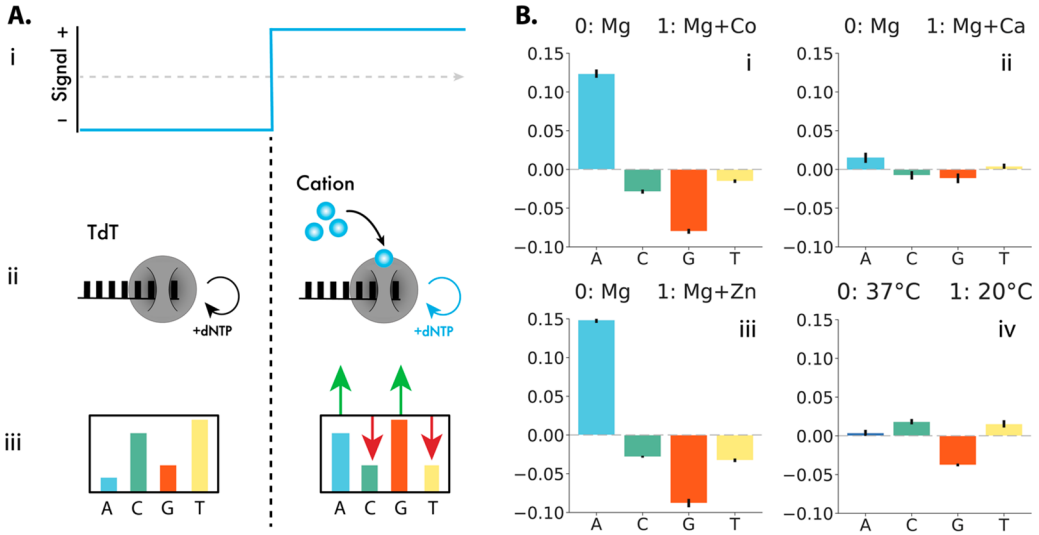


Figure 1. TURTLES device architecture and environmental signal responses. (A-i) A representative time-varying input signal. (A-ii) TdT interacts directly with the signal of interest (small blue circles). (A-iii) This results in different average DNA compositions under each condition. (B) Change in frequency of nucleotide selectivity by TdT in the presence of various environmental signals tested. Signal 0 was 10 mM Mg^{2+} at 37 °C for 1 h. Signal 1 was (i) 10 mM Mg^{2+} + 0.25 mM Co^{2+} at 37 °C for 1 h, A incorporation increased by 12.4%, while G decreased by 8.0%, and T and C decreased by 1.5% and 2.9%, respectively, (ii) 10 mM Mg^{2+} + 1 mM Ca^{2+} at 37 °C for 1 h, A increased by 1.5%, G decreased by 1.2%, T increased by 0.4%, and C decreased by 0.8%; (iii) 10 mM Mg + 20 μM Zn^{2+} at 37 °C for 1 h, A increased by 14.9%, G decreased by 8.8%, T decreased by 3.3%, and C decreased by 2.8%, and (iv) 10 mM Mg^{2+} at 20 °C for 1 h, 0.4% increase in A, 3.8% decrease in G, 1.5% increase in T, and 1.8% increase in incorporation of C. Error bars show two standard deviations of the mean. The statistical significance was assessed after first transforming the data into Aitchison space, which makes each dNTP frequency change statistically independent of the others (Figure S2).

ration occurs on a time scale on the order of milliseconds to seconds,[15] potentially enabling orders of magnitude improvements in the temporal accuracy and resolution of molecular recording. However, prototypical DNAps replicate the contents of an existing strand, which would prevent recording of new information. A DNAp that does not simply replicate DNA but rather creates a *de novo* sequence could allow for DNA recording.

Terminal deoxynucleotidyl transferase (TdT) is a DNAp that can randomly incorporate bases to the 3′-position of a DNA strand with biases toward particular bases.[16,17] Shifting the nucleotide bias of TdT could make it a prime candidate for post-translational control of DNA encoding. In fact, *in vitro* experiments have shown that cations (e.g., Co^{2+}) can shift the bias of TdT.[16,18] In addition, DNA is synthesized in a sequential manner, which provides an estimate of the time a particular base is added. We therefore hypothesized that the environment in which a TdT extends a DNA strand might be encoded by the average base composition of the extended DNA. Put another way, by a combination of the change in nucleotide bias in the presence of cations and the addition of time bases inferred from sequence, a molecular ticker tape may be possible.[13,14]

Here, we introduce TdT-based Untemplated Recording of Temporal Local Environmental Signals (TURTLES), a polymerase-based molecular recording system that achieves high time resolution *in vitro* by utilizing post-translational control to change the bases incorporated. First, we describe methods to characterize DNA sequences synthesized by TdT and show that cation concentrations can be encoded in populations of TdT-synthesized DNA using an approach that analyzes the average composition of several bases added at similar times on the same or parallel strands of DNA. We next developed an algorithm to accurately estimate the times of signal changes and show that temporal information can be accurately recovered by using estimates of DNA synthesis rates to map DNA sequences to real time. We also describe an expanded TURTLES system that uses an engineered, allosterically modulated TdT to expand the generalizability and tunability of the system. By inserting an exogenous sensing domain, we show that TURTLES can be adapted to the arbitrary signals of interest. When they are taken together, these results establish the feasibility of DNA synthesis-based encoding systems and demonstrate a recording of cationic environmental signals with minutes resolution for enhanced applications in DNA data storage and DNA recording.

RESULTS

TdT Can Encode Environmental Signals *In Vitro* via Changes in Base Selectivity. The cations present in the reaction environment of TdT affect the rate of incorporation for specific nucleotides.[16] For example, previous studies[18−20] and our experiments show that, when only one nucleotide is present, the incorporation rates of pyrimidines, dCTP and dTTP, increase in the presence of Co^{2+} (Figure S1).

We sought to examine if these Co^{2+}-dependent changes in kinetics also occurred in the presence of all four nucleotides, dATP, dCTP, dGTP, and dTTP (hereafter referred to as A, C, G, and T). The nucleotide composition of ssDNA extended by bovine TdT in a cobalt-free reaction buffer or with cobalt added was determined by next-generation sequencing. In the presence of Co^{2+}, A incorporation increased, while G, T, and C incorporation decreased (Figure 1B-i and Figure S2). Notably, the significant difference in the composition of DNA under each condition effectively encodes information about the environmental Co^{2+} concentration at the time of DNA synthesis.

Next, we were interested in understanding which conditions could be encoded by TURTLES. We examined Ca^{2+}, Zn^{2+}, and temperature. Ca^{2+} signaling is biologically ubiquitous and

https://doi.org/10.1021/jacs.1c07331
J. Am. Chem. Soc. 2021, 143, 16630−16640

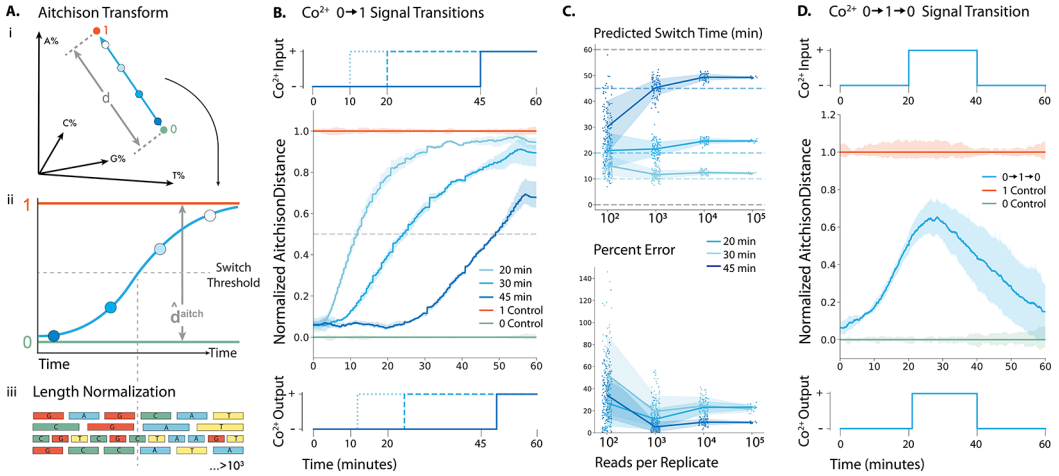


Figure 2. Recording Co^{2+} fluctuations into ssDNA with minutes resolution *in vitro*. (A-i) Representation of how the percent incorporation of each nucleotide is dependent on each of the nucleotides incorporated. (A-ii) Sequences were normalized by length before the nucleotide composition at each time point was calculated. (A-iii.) By transforming the percent incorporation of each nucleotide to the Atichison distance, we can calculate the total "output signal". We plot the Aitchison distance for the recording experiment between the 0 (green) and 1 (orange) signals. (B) (top) 0.25 mM Co^{2+} was added at 10, 20, and 40 min to generate a 0 → 1 transition. (center) Mean output signal across three biological replicates. Vertical lines are drawn at the inferred transition time. (bottom) Predicted output signal transition times were 11.9, 24.4, and 49.2 min. (C) (top) Predicted switch times for each 0 → 1 transition calculated from randomly sampled subsets of sequences. (bottom) Time prediction error for each 0 → 1 transition calculated from randomly sampled subsets of sequences. (D) (top) 0.25 mM Co^{2+} was added at 20 min and then removed at 40 min to generate a 0 → 1 → 0 transition. (center) Mean output signal across three biological replicates. (bottom) Using the algorithm detailed by Glaser et al.,[13] the signal was deconvoluted into a binary response, with vertical lines drawn at the predicted switch times of 21 and 41 min.

functions in neural firing,[21] fertilization,[22,23] and neurodevelopment,[24] Zn^{2+} is an important signal in the development and differentiation of cells,[25] and the temperature is relevant in many situations.

Each signal altered both the particular dNTPs affected and the magnitude of the change in dNTP selectivity. We were able to encode 1 mM Ca^{2+}, 20 μM Zn^{2+}, and a temperature of 20 °C (Figure 1B-ii–iv and Figure S2). Both cation addition and a temperature change also altered the lengths of ssDNA strands synthesized (Figures S3–S8). For each environmental condition tested, we observed significant differences in the composition of TdT-synthesized DNA. We conclude that input-dependent changes in TdT nucleotide selectivity can encode environmental information into DNA. For further analysis we chose to focus on Co^{2+} as the candidate cationic signal due to the large difference in TdT selectivity.

Recording a Single Step Change in Co^{2+} Concentration with Resolution of Minutes. Having shown nucleotide selectivity changes in the presence of Co^{2+}, we attempted to identify the time at which Co^{2+} was added to a TdT-catalyzed ssDNA synthesis reaction on the basis of the changes in the nucleotide composition of the synthesized ssDNA (Figure 2A). During a 60 min extension reaction, we created step transitions in cobalt concentration by adding 0.25 mM Co^{2+} at 10, 20, and 45 min (hereafter referred to as a 0 → 1 input, where 0 is Co^{2+} free and 1 is with 0.25 mM added Co^{2+}) (Figure 2B, top). For each reaction, we analyzed approximately 500000 DNA strands by deep sequencing and calculated the dNTP incorporation frequencies over all reads. Because the change in dNTP selectivity is a compositional data type (i.e., all changes in base frequency sum to 0%), they are not independent and do not satisfy the independence assumption required for most statistical tests. Therefore, to perform hypothesis testing, the base composition was transformed into Aitchison space, where the proportion of each base becomes independent of the other three. The output signal for each reaction was calculated as the normalized distance in Aitchison space between the 0 and 1 controls.

After each sequence was normalized by its own length, the Aitchison location along the extended strands showed that later addition of Co^{2+} resulted in dNTP selectivity changes farther down the extended strand (Figure 2B, center). To estimate the real time at which changes occurred, we calculated the average location along the strand for all sequences under each condition at which a distance halfway between the 0 and 1 control output signals was reached. To translate this location into a particular time in the experiment, we calculated the average rate of dNTP addition in each state (Figure S9) and derived an equation that adjusted for the change in rate of DNA synthesis between the 0 and 1 controls (eq 5 and Supplementary Methods 3). Using this information, we estimated that the Co^{2+} additions were made at 11.9, 24.4, and 49.2 min (Figure 2B, bottom). We were also able to estimate the time within 4 min of the unit input step function for the reverse, a 1 → 0 condition (Figure S10).

While we were able to accurately estimate the times of Co^{2+} addition (0 → 1) and removal (1 → 0), simultaneously synthesizing ~500000 strands of DNA is infeasible for certain applications. To determine the number of strands needed for a reasonable statistical certainty, we randomly sampled smaller groups of strands from the experiment and evaluated our ability to predict when Co^{2+} was added (Figure 2C, top and bottom). With about 1000 strands, we still estimated the time of Co^{2+} additions to within 2 min of actual input times (Figure 2C). Thus, TURTLES can be employed for recording temporal information even with a limited number of ssDNA substrates.

Recording Multiple Fluctuations in Co^{2+} Concentration onto DNA with a Resolution of Minutes. In contrast to current DNA-based recorders, which rely on time-integrated recording methods (i.e., accumulation of mutations) or slow signal transducing steps, DNA synthesis-based

https://doi.org/10.1021/jacs.1c07331
J. Am. Chem. Soc. 2021, 143, 16630–16640

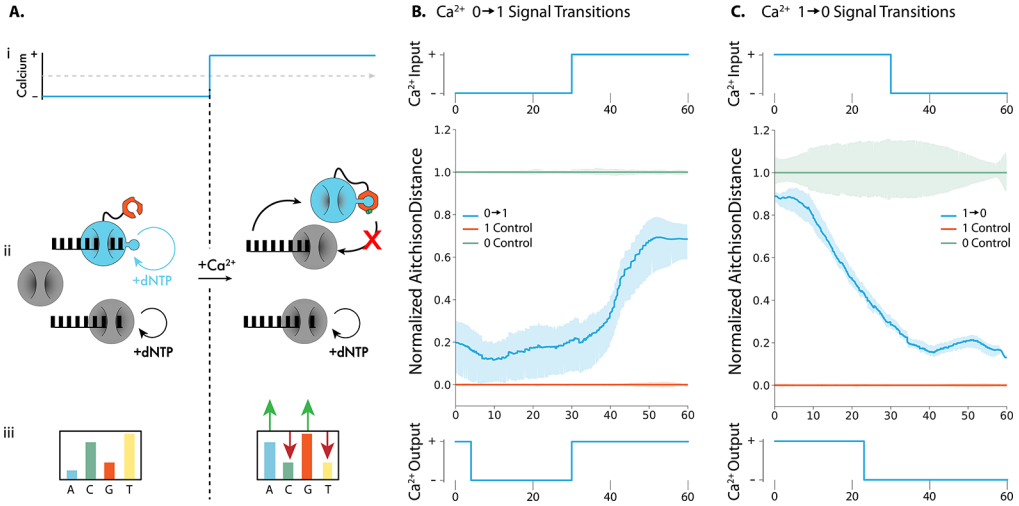


Figure 3. Recording Ca^{2+} changes into ssDNA with TURTLES-2. (A.i.) Representative calcium transition changing concentration from calcium-free to calcium-added conditions during a TdT-based DNA synthesis reaction. Mg^{2+} concentration and reaction temperature are held constant. (A.ii.) The CaM subunit (orange) of engineered TdT (teal) binds with the fused M13 peptide in the presence of Ca2+ to allosterically turn off DNA synthesis. The activity of the reference TdT (gray) is not affected by Ca^{2+}. In the absence of Ca^{2+} both the engineered TdT and reference TdT are carrying out DNA synthesis. In the presence of Ca^{2+} only the reference TdT synthesizes DNA. (A.iii.) This results in a change in the overall nucleotide incorporation preference upon a change in Ca^{2+}. (B) Top: 100 μM $CaCl_2$ was added to the extension reaction at 30 min to generate a 0 → 1 transition. Center: Mean output signal across three biological replicates. Bottom: Using a modified version of the algorithm detailed in Glaser et al.,[13] the signal was deconvoluted into a binary response, with the predicted switch time of 30 min. (C) Top: 50 μM EGTA was added to the extension reaction at 30 min to generate a 1 → 0 transition. Center: Mean output signal across three biological replicates. Bottom: Using a modified version of the algorithm detailed in Glaser et al.,[13] the signal was deconvoluted into a binary response, with the predicted switch time of 22 min.

approaches can record the dynamics of multiple fluctuations in real time. While accumulation can tell what fraction of the time a signal was present in a time period, the ability to record multiple temporal changes would enable new levels of insight into dynamic processes such as physiological signaling, which are poorly captured by time-integrated recording methods.

We used TURTLES to record a 0 → 1 → 0 input cobalt signal. The 0 condition was maintained for the first 20 min, 1 for the next 20 min, and 0 for the last 20 min of the extension reaction (Figure 2D, top). Using the same methods as for the single-step transition, we calculated the output signal (Figure 2D, center). To account for the additional complexity of multiple fluctuations, we used an algorithm previously developed by Glaser et al.[13] (see Materials and Methods for details) to binarize the value of the output signal every 0.1 min. We were able to accurately reconstruct the input 0 → 1 → 0 signal, estimating transitions between the 0 and 1 signals occurring at 21 and 41 min on the basis of sequencing data (Figure 2D, bottom).

On the basis of the measured experimental parameters, we used *in silico* simulations to estimate the performance of TURTLES in more complex recording environments. We investigated how rapidly signals could change and still be detected and how many consecutive condition changes could be recorded accurately. By varying the length of time of each input condition (0 or 1) from 1 to 20 min (Figure S12A), we estimated that TURTLES can record 6 consecutive signal changes with 1 min between each with >75% accuracy from ≥2000 strands of 100bp ssDNA synthesized (>90% accuracy with ≥60000 strands of ssDNA of 50 bp length each) (Figure S12B). By keeping the duration of each input condition (0 or 1) constant at 10 min (Figure.S12C) and varying the total number of condition changes, we estimated that TURTLES would be capable of recording 10 sequential input signal changes with >80% accuracy (Figure S12D). We thus show that TURTLES has the potential for high temporal precision and can decode signals across a range of frequencies.

TdT Can Be Engineered to Allosterically Respond to and Encode Environmental Signals. Unlike Co^{2+} and Zn^{2+}, we observed that Ca^{2+} only modestly altered the dNTP selectivity of TdT, precluding temporal recordings of Ca^{2+} concentration. To show that TURTLES could be expanded to signals to which TdT was unresponsive or weakly responsive, we attempted to engineer a TdT to allosterically respond to Ca^{2+}. The structural determinants of base selectivity in TdT are poorly understood, which ruled out directly increasing the dynamic range of Ca^{2+}-responsive dNTP selectivity changes.[26] Accordingly, we conceived a modular recording system based on two distinct TdT species with different inherent dNTP selectivity.

The two-TdT system, TURTLES-2, uses a reference TdT whose catalytic rate is unaffected by inputs and a sensor TdT that is allosterically activated or deactivated in response to input signals. By the choice of a pair of sensor and reference TdTs with distinct nucleotide selectivity, TURTLES-2 encodes environmental signals into changes in DNA composition on the basis of the differential activity of the two TdTs (Figure 3A). As the sensing and recording functions of the system are distributed between two TdT variants, TURTLES-2 is more accessible to tuning and engineering efforts.

We employed the natural calcium sensing protein calmodulin (CaM) and the cognate binding peptide M13 to generate a TdT with allosterically modulated activity. The calcium-dependent interaction between CaM and M13 has been previously utilized in the engineering of allosteric calcium biosensors[27,28] and as a platform for generalizable ligand biosensors.[29,30]

https://doi.org/10.1021/jacs.1c07331
J. Am. Chem. Soc. 2021, 143, 16630−16640

Here, we generated variants with M13 fused to one of four sites in mTdT that were predicted to minimize the structural disruption of inserting the M13 sequence using SCHEMA-RASPP[31] (Figure S13 and Table S2). After an initial activity screening (Figure S14) of the variants, we observed that one variant, mTdT(M13-388), retained polymerase activity. CaM was subsequently fused to the N-terminus of mTdT(M13-388) via a linker. Primer extension reactions showed that the resulting CaM-mTdT(m13-388) variant was active under calcium-free conditions and inactive under calcium-added conditions (Figure S15). To confirm that the calcium-dependent interaction between CaM and M13 was responsible for the observed activity modulation, we mutated four essential Ca^{2+}-binding residues in CaM, which ablated the calcium sensitivity of CaM-mTdT(M13-388) (Figure S15).

Depending on the application, sensor polymerases with different calcium affinities may be useful to selectively record Ca^{2+} fluctuations exceeding threshold concentrations. We anticipated that the modular design of CaM-mTdT(M13-388) would allow the properties of the fusion to be rationally modified with CaM variants with known differences in Ca^{2+} affinity. The polymerase activity was determined by the length distributions of primer extensions, CaM-mTdT(M13-388) variants containing the CaM mutants D96V, D130G, and D142L, which reduce the calcium affinity of CaM.[32] We observed that all variants exhibited greater activity in comparison to the unmodified CaM-mTdT(m13-388) in the presence of low concentrations of calcium (Figure S16). Moreover, the increase in activity correlated with the reported effective Ca^{2+} K_D value of the variants, demonstrating that the calcium sensitivity of CaM-mTdT(M13-388) can be rationally tuned.

Next, we tested if the TURTLES-2 system could encode the Ca^{2+} state into DNA. CaM-mTdT(M13-388) was purified, and an NGS analysis of extension reactions performed with the polymerase confirmed the calcium-sensitive phenotype (Figures S17 and S18). We characterized CaM-mTdT(M13-388) in the context of the TURTLES-2 recording system by performing extensions with a mixture of purified bovine TdT and CaM-mTdT(M13-388) under calcium-free and calcium-added conditions. The two-polymerase system exhibited a significantly altered nucleotide selectivity under the calcium-added conditions (Figure S19). As expected, under the calcium-free conditions the overall base incorporation preference was approximately the average of the observed preferences of bovine TdT and CaM-mTdT(M13-388), whereas the overall base preference under the calcium-added conditions was nearly identical with that of bovine TdT (Figure S20). We conclude that the differential overall base selectivity of the TURTLES-2 system is capable of encoding the environmental calcium state into DNA.

Recording a Single-Step Change in Ca^{2+} Concentration with a Resolution of Minutes with Two TdT Systems. We next investigated if the differential calcium response of TURTLES-2 could be used to infer the time at which calcium concentrations changed in an extension reaction. During a 60 min extension we tested both 0 → 1 and 1 → 0 step transitions at 30 min, where 0 is calcium-free and 1 is calcium-added (Figure 3B, top, Figure 3C, top, Supplementary Note 6). Using a variation of the model developed by Glaser et al.,[13] we inferred transition times of 22 min for the 1 → 0 transition (Figure 3B, bottom) and 30 min for the 0 → 1 transition (Figure 3C bottom). The decreased accuracy of the estimated time for the calcium transitions did not correspond to an increase in measurement variability. We speculate that the offset may be due to the different kinetics of *holo* calmodulin binding M13 and *apo* calmodulin releasing M13. Time estimations for transitions in TURTLES-2 would likely improve with more sophisticated decoding algorithms and a deeper characterization of the transition behavior of CaM-mTdT(M13-388). We conclude that TURTLES-2 serves as a promising proof of concept for developing high-resolution temporal calcium signal encoding systems.

DISCUSSION

While many DNA-based biosensors have been deployed for studying physiological signals of interest,[33−37] the scalability and spatial resolution of biosensors are intrinsically limiting in some applications.[38] By leveraging the *post hoc* recovery of biological data, optimized TURTLES systems may be capable of enabling otherwise inaccessible high-resolution spatial and temporal recordings of physiological signaling molecules that fluctuate on the time scale of 10^1−10^3 minutes. Such signals include slow calcium signaling that occurs in neurons[13,14,38,39] and vertebrate development.[22] Additional optimization of the TURTLES system will be required to enable the spatiotemporal resolution for characterizing systems with shorter time scales.

Beyond biological applications, there has been a sustained interest in biosensors for testing environmental parameters such as water quality. For longer-term tracking of contaminating metal ions in water, one could use TURTLES to track the cobalt concentration over time.[40,41] In concert with microfluidic reaction control, TURTLES can also serve as a competitive platform for enzymatic DNA synthesis for data archiving,[9−11] which is an appealing alternative to phosphoramidite methods due to the low cost and reduced environmental impact.[42] Although TURTLES was used in this work to record binary signals, it could also be used to record more than two signals in sequence. For example, *in vitro* addition of nucleotides via TdT with varied input nucleotide compositions could be used to encode data (Figure S21). While the information density of such a system would be lower than those using base-specific DNA synthesis, it would not require specialized substrates or complex reaction cycling and would thus be a competitive application for digital data storage.

Going forward, more sophisticated computational methods will improve the recording accuracy of TURTLES. In this study we utilized simple, intuitive models of TdT activity to transform sequence data into temporal information. By incorporating kinetic models of TdT activity or machine learning to classify signal changes along individual DNA strands, the accuracy of temporal estimations could likely be increased. These methods would also improve the robustness of TURTLES recordings by reducing the required sequencing depth from thousands to hundreds of reads. In this work, both the inputs and outputs for TURTLES were binarized; however, the underlying principle can be extended to record continuously varying analog signals with improved decoding algorithms.

The quality of TURTLES recordings may also be improved by engineering the properties of TdTs. In particular, the sequencing depth required to accurately decode recordings can be reduced by increasing the magnitude of changes in nucleotide selectivity in response to inputs. Likewise, reference TdTs that have a more distinct nucleotide selectivity from the

https://doi.org/10.1021/jacs.1c07331
J. Am. Chem. Soc. 2021, 143, 16630−16640

CaM-mTdT(M13-388) sensor TdT can be engineered or identified among natural TdT diversity. Improvements to the temporal resolution of TURTLES systems can be accomplished by enhancing the nucleotide incorporation rate of TdT. In TURTLES-2, the structural optimization of CaM-mTdT-(M13-388) may improve temporal resolution by optimizing the kinetics of the CaM−M13 interaction. Notably, fluorescent biosensors based on CaM−M13 interactions can report calcium spikes on the order of seconds,[43,44] suggesting that calcium sensing will not be limiting with respect to temporal resolution in an optimized system. The functionality of TURTLES-2 may be further expanded by employing generalizable sensors based on the CaM−M13 interaction[29] or by probing TdT with sensing domains other than calmodulin such that new signals of interest can be encoded or recorded. In all, we have demonstrated a new methodology for recording dynamic, environmental information into DNA that relies only on allosteric regulation, enabling a resolution of minutes.

CONCLUSION

In this study, we demonstrated two DNA synthesis-based recording concepts that encode and record the temporal dynamics of fluctuating environmental signals with an accuracy of minutes. By coupling sensing and writing functions, TURTLES simplifies the recording apparatus to a post-translational system. This gives TURTLES distinct advantages over the temporal constraints of existing tools, paving the way toward the development of tools with heretofore unprecedented temporal accuracy and resolution. While TURTLES can record several physiologically relevant signals, TURTLES-2 lends tunability to the recording system with simple rational engineering. Given the uncomplicated and genetically encodable design of TURTLES systems, we anticipate that TURTLES can be further developed for both *in vitro* and *in vivo* biorecording applications.

MATERIALS AND METHODS

Enzymes and ssDNA Substrate. Terminal deoxynucleotidyl polymerase, T4 RNA ligase I, and Phusion High-Fidelity PCR Master Mix with HF Buffer were purchased from New England Biolabs (NEB). ssDNA substrates used for extension reactions were purchased from Integrated DNA Technologies (IDT) with standard desalting. dNTPs were obtained from Bioline.

CaM Fusion Design and Screening. Four fusion proteins were designed that consisted of CaM fused to the N terminus of mTdT by a $(GGGGS)_4$ linker and M13 inserted immediately following the fusion residue (see below) with flanking GS linkers. Fusion sites were selected from crossover sites identified with the SCHEMA/RASPP algorithm, on the basis of which sites were in catalytically essential regions and would be sterically available to CaM. SCHEMA crossover sites were calculated according to previously described protocols.[31] Sets of crossover points were calculated for 3, 4, 5, 6, and 7 total crossovers. Calculations were performed with the following parameters: minimum fragment length 4, bin width 1, parent sequence NP_001036693.1, parent structure PDB 4I27 (all ligands, metals, and waters removed), homology sequences NP_803461.1, AAH12920.1, NP_001012479.1, XP_021064401.1, and XP_020136193.1. All sequences were trimmed to only include residues crystallized in the parent structure. Fusion sites were selected from crossover points that were in the DNA-binding region of mTdT (residues 282, 284, and 287) or in Loop 1, a catalytically essential structure (residue 388). M13 fusions were screened for activity without N-terminal CaM to validate that the fusion was tolerated.

Cloning CaM-TdT(M13) Variants. Molecular cloning of DNA constructs was completed under a contract research agreement with the laboratory of Dr. J. Andrew Jones at Miami University, Oxford, OH. The pET28a-M-CaM-cTdTS-M13-XXX (282, 284, 287, and 388) variants were constructed using a two-part Gibson assembly method. The approximately 75bp M13 fragment was amplified from linear double-stranded DNA template (gBlock-CaM-Linker-M13, IDT) using Accuzyme DNA polymerase (Bioline) using DNA primers P2 P28 given in Table S1. The amplicon was then purified using a Cycle Pure Kit (Omega Biotek). The vector backbone fragment was amplified from pET28a-M-CaM-cTdTS plasmid DNA constructed above using *PfuUltra* II Hotstart PCR Master Mix using DNA primers P29−P36 given in Table S1. The PCR product was then digested with *Dpn*I to remove the DNA template. The approximately 8100bp amplicon was purified using a gel extraction kit (Omega Biotek). DNA concentrations of both linear fragments were measured using a Take3 plate coupled with a Biotek Cytation 5 plate reader. The corresponding backbone and M13 fragments were then assembled using the repliQa HiFi Assembly Kit (Quanta bio), transformed into chemically competent DH5α, and selected on LB-Kanamycin (50 μg/mL) plates. Individual colonies were then screened via restriction digestion and verified using Sanger sequencing (CBFG, Miami University) with primers S1−S8 (Table S1).

CaM-TdT(M13-388) Expression and Purification. Purification optimizations determined that N-terminal MBP was unnecessary for expression and purification and was not included in the final expression construct. The expression construct (pET28a-CaM-mTdT(m13-388)) was transformed into chemically competent NEB T7Express cells, plated on kanamycin-selective plates, and incubated at 37 °C. The following day, a single colony was selected and inoculated into 5 mL of kanamycin-supplemented LB. The culture was incubated for 20 h at 37 °C. Four flasks with 120 mL of kanamycin-supplemented LB were inoculated 1/400 (v/v) with the overnight culture. The cultures were incubated with shaking at 250 rpm. Once the OD_{600} value was between 0.5 and 0.6, the cultures were cooled to room temperature and induced with 1 mM IPTG. Following induction, the cultures were incubated for 18 h at 15 °C. The cells were pelleted at 4 °C, and the supernatant was discarded. The decanted cell pellets were stored at −80 °C.

The cell pellets were thawed on ice. Lysis and affinity chromatography were performed using the Takara Bio HisTALON gravity column purification kit; all steps were performed according to the manufacturer's native protein extraction protocol. Note that the cell pellets were treated with optional DnaseI and lysozyme during lysis. A 1 mL portion of Takara Bio TALON metal affinity resin was used for affinity chromatography. All binding and washing steps were performed on ice with shaking at 250 rpm. Fifteen bed volumes of wash buffer were used for all washes. CaM-mTdT(M13-388) was eluted from the resin in 10 500 μL fractions. Each fraction was analyzed by SDS-PAGE, and the total protein concentration in each fraction was measured by absorbance at 280 nm. The first five elution fractions, which contained the majority of eluted protein, were pooled.

The pooled fractions were diluted in binding buffer (20 mM Tris-HCl, pH 8.3) and further purified by anion exchange chromatography using a Cytiva HiTrap Q HP 5 mL column and a 40 CV gradient from 0 to 1 mM NaCl in binding buffer with a GE Healthcare AKTAxpress FPLC apparatus. The protein eluted in two fractions.

Both elutions were buffer-exchanged by dialysis into a storage buffer consisting of 200 mM KH_2PO_4 and 100 mM NaCl at pH 6.5 and concentrated using Vivaspin 20 columns to a final concentration of 0.37 mg/mL for the first fraction and 0.98 mg/mL for the second fraction. The fractions were aliquoted and flash frozen on dry ice for storage at −80 °C. Notably, a PAGE analysis showed that the second elution contained a product at 25 kDa in addition to the expected fusion protein at approximately 70 kDa. Both CaM-mTdT(M13-388) elutions recapitulated the calcium-sensitive phenotype and exhibited similar nucleotide selectivities (Figures S18 and S19). As significantly more protein was recovered in the second elution, it was used for all subsequent experiments.

Cell-Free Protein Expression and Primer Extension Assay. For initial activity screening of fusion variants and CaM-mTdT(M13-

https://doi.org/10.1021/jacs.1c07331
J. Am. Chem. Soc. 2021, 143, 16630−16640

388) characterization, proteins were expressed in cell-free reactions. Variants were expressed using NEB PURExpress in 25 μL reactions containing 40% (v/v) PURExpress Solution A, 30% (v/v) PURExpress Solution B, 1.6 U/μL NEB Rnase I, 10 ng/μL expression vector DNA, and dH_2O to volume. The expression reaction mixtures were incubated for 4 h at 30 °C.

Primer extension reactions were prepared on ice. Primer extensions were performed in 25 μL reactions containing 1X NEB TdT Reaction buffer, 0.8 μM single-stranded, FAM-labeled substrate DNA FAM_NB (Table S2), 1 mM dNTPs, polymerase, and dH_2O to 25 μL. For variants expressed in PURExpress, 2.5 μL of the expression reaction was used immediately after expression, 20U (approximately 0.2 μg) of the NEB bovine TdT was used for positive control reactions, and approximately 0.5 μg of purified CaM-mTdT(M13-388) was used for activity validation reactions after purification. For calcium-added conditions, $CaCl_2$ was added to the reaction mixtures to a final concentration of 1 mM. Extension reaction mixtures were incubated for 2 h at 37 °C.

Completed extensions were analyzed by urea-PAGE. A 8 μL portion of each completed primer extension reaction was combined with 12 μL of BioRad 2x TBE urea sample buffer and boiled for 10 min. The boiled samples were loaded onto a 10% polyacrylamide TBE urea gel (bioRad 4566036), and 200V was applied to the gel for 40 min. The gels were imaged on a GE Healthcare Typhoon 9400 laser scanner using a 200 μm pixel size and λ_{ex} = 488 nm and λ_{em} = 520 nm BP40. The imaging gain was adjusted for each experiment to avoid saturation.

Extension Reaction for Calculating the Effects of Co^{2+}, Ca^{2+}, Zn^{2+}, and Temperature on the Overall dNTP Preference of TdT. Each extension reaction consisted of a final concentration of 10 μM ssDNA substrate CS1 (Table S2), 1 mM dNTP mix (each dNTP at 1 mM final concentration), 1.4x NEB TdT reaction buffer, and 10 units of TdT to a final volume of 50 μL. When the effect of cations was tested, $CoCl_2$ was added at a final concentration of 0.25 mM, $CaCl_2$ at 2 mM, or $Zn(Ac)_2$ at 20 μM. It is important to note that reaction initiation was done by adding TdT to the ssDNA substrate mix (ssDNA substrate mix consisted of the ssDNA substrate, dNTPs, and the cation). Prior to reaction initiation, the ssDNA substrate mix and TdT were stored in separate PCR strip tubes at 0 °C (on ice). The reaction was run for 1 h at 37 °C in a Bio-Rad PCR block. When the effect of temperature was tested, the same reaction mix was run on a Bio-Rad PCR block set at the tested temperatures for 1 h. Reactions were stopped by freezing at −20 °C. For initial testing, reactions were analyzed by urea-PAGE; 2 μL of the reaction mixture was mixed with 12 μL of TBE-Urea (Bio-Rad) loading dye and boiled for 10 min at 100 °C. All of the diluted extension reaction mixture was then loaded onto 30 μL, 10-well 10% TBE-Urea Gel (Bio-Rad) and run for 40 min at 200V. Immediately after the run was over, the gel was stained with Sybr Gold for 15 min and imaged on an ImageQuant BioRad instrument.

TURTLES 0 → 1 Extension Reactions. Mg^{2+} only for 1 h (signal 0) and Mg^{2+} + Co^{2+} for 1 h (signal 1) were set up as the regular extension reactions mentioned above. The 0 → 1 reactions where the signal changed from 0 to 1 at various times during the 1 h extension were run starting at a total volume of 45 μL with Mg^{2+} only. A 5 μL portion of 2.5 mM $CoCl_2$ was added at the time we wanted the signal to change from 0 to 1. Reactions were all run for a total of 1 h in triplicate. Fresh signal 0 and signal 1 controls were run with each setup.

TURTLES-2 Controls and 0 → 1 and 1 → 0 Extension Reactions. TURTLES-2 extension reaction mixtures contained 1X NEB TdT reaction buffer, 0.1 μM ssDNA substrate CS1_5N (Table S2), 1 mM dNTP mixture, and 2.5 μL of a polymerase mixture. The polymerase mixture contained CaM-mTdT(M13-388) at a concentration of 0.45 mg/mL and NEB TdT at a concentration of 0.002 mg/mL (0.4U). Calcium-free reactions included EGTA at a final concentration of 50 μM. The high calcium control for 0 → 1 reactions was supplemented with $CaCl_2$ and EGTA to final concentrations of 100 and 50 μM, respectively. The high calcium control for 1 → 0 reactions was not supplemented with EGTA or $CaCl_2$ (Supplementary Note 6). Reaction mixtures were brought to a final volume of 25 μL with nuclease-free water. Reaction mixtures were assembled on ice and initiated by adding TdT to the substrate mixture. Reaction mixtures were incubated for 1 h at 37 °C in a Bio-Rad PCR block and terminated by heating to 80 °C for 10 min.

Signal transitions were performed by 1 μL additions at 30 min for 1 → 0 reactions; the addition contained 1X NEB TdT buffer and 1.3 mM EGTA (50 mM EGTA in final reaction postaddition). For 0 → 1 reactions, the addition contained 1X NEB TdT buffer and 2.6 mM $CaCl_2$ (100 μM $CaCl_2$ in final reaction postaddition).

Extension Reactions for 0 → 1 → 0 Setup. Mg^{2+} only for 1 h (signal 0) and Mg^{2+} + Co^{2+} for 1 h (signal 1) were set up as the regular extension reactions mentioned above. The 0 → 1 → 0 reactions where the signal changed from 0 to 1 at 20 min and back to 0 at 40 min were run starting at a total volume of 45 μL with Mg^{2+} only. A 5 μL portion of 2.5 mM $CoCl_2$ was added at the time we wanted the signal to change from 0 to 1. To change the signal from 1 to 0, since the ssDNA was suspended in reaction buffer for these setups, we used a ssDNA cleanup kit (methods mentioned below) to remove the reaction buffer, TdT, cation, and dNTPs from each reaction mixture. All of the ssDNA collected from the ssDNA cleanup kit (20 μL) was then prepared for the last part of the extension reaction. Collected ssDNA was mixed with a dNTP mix at a final concentration of 1 mM (each dNTP at 1 mM final concentration), 1.4x TdT reaction buffer, and 10 units of TdT to a final volume of 50 μL. All reactions were always initiated by adding TdT in the end. Signal 0 and signal 1 controls were run for 1 h for each setup in triplicate and also put through the ssDNA wash step at 40 min. Six replicates were run for 0 → 1 → 0 reactions.

ssDNA Wash for Replacing Buffers for 0 → 1 → 0 Reactions. To change the cation concentration from 1 to 0, we utilized the ssDNA cleanup kit (ssDNA/RNA clean/concentrator D7010) from Zymo Research such that all of the extended ssDNA synthesized in the initial part of the experiment was retained on the column and the TdT, reaction buffer, cation, and dNTPs were washed away. Each 50 μL extension reaction mixture was individually loaded into a separate column. Protocol was followed as mentioned in the kit. ssDNA was eluted into 20 μL of ddH_2O. We noticed in initial tests that, after the ssDNA cleanup kit was used, there was little to no TdT-based extension in some replicates (data not included). We presume this is due to some ethanol getting carried forward into the eluted ssDNA. Thus, we extended the dry spin time on the basis of a suggestion from Zymo Research to 4 min. We also utilized two other ways to evaporate any remaining ethanol after the column dry spin step based on the protocol mentioned in Cold Spring Harbor Protocols.[45] We either kept the columns open in a biohood for 15 min to allow for evaporation or after elution of ssDNA we kept the 1.5 mL Eppendorf tubes containing the eluted ssDNA open at 45 °C for 3 min. Both methods gave better ethanol removal than just dry spin, and they were tried in triplicate and averaged and plotted for the time prediction analysis (Figure 2D).

Illumina Library Preparation and Sequencing. Our sample preparation pipeline for NGS was adapted from a previous protocol.[46,47] After an extension reaction, 2 μL of the product was utilized for a ligation reaction. A 22bp universal tag, common sequence 2 (CS2) of the Fluidigm Access Array Barcode Library for Illumina Sequencers (Fluidigm), synthesized as ssDNA with a 5′-phosphate modification and PAGE purified (Integrated DNA Technologies), was blunt-end ligated to the 3′-end of extended products using T4 RNA ligase. Ligation reactions were carried out in 20 μL volumes and consisted of 2 μL of the extension reaction mixture, 1 μM of CS1 ssDNA, 1X T4 RNA Ligase Reaction Buffer (NEB), and 10 units of T4 RNA Ligase 1 (NEB). Ligation reaction mixtures were incubated at 25 °C for 16 h. Ligated products were stored at −20 °C until PCR, which was carried out on the same day. Ligation products were never stored at −20 °C for more than 24 h.

PCR was performed with barcoded primer sets from the Access Array Barcode Library for Illumina Sequencers (Fluidigm) to label extension products from up to 96 individual reactions. Each PCR primer set contained a unique barcode in the reverse primer. From 5′-

https://doi.org/10.1021/jacs.1c07331
J. Am. Chem. Soc. 2021, 143, 16630−16640

3′ the forward PCR primer (PE1 CS1) contained a 25-base paired-end Illumina adapter 1 sequence followed by CS1. The binding target of the forward PCR primer was the reverse complement of the CS1 tag that was used as the starting DNA substrate. From 5′-3′ the reverse PCR primer (PE2 BC CS2) consisted of a 24-base paired-end Illumina adapter 2 sequence (PE2), a 10-base Fluidigm barcode (BC), and the reverse complement of CS2. CS2 DNA that had been ligated onto the 3′-end of extended products served as the reverse PCR primer-binding site. Each PCR reaction consisted of 2 μL of ligation product, 1X Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB), and 400 nM forward and reverse Fluidigm PCR primers in a 20 μL reaction volume. Products were initially denatured for 30 s at 98 °C, followed by 20 cycles of 10 s at 98 °C (denaturation), 30 s at 60 °C (annealing), and 30 s at 72 °C (extension). Final extensions were performed at 72 °C for 10 min. Amplified products were stored at −20 °C until clean up and pooling. QC for individual sequencing libraries was performed as follows. A 2 μL portion of each library was pooled into a QC pool, and the size and approximate concentration were determined using an Agilent 4200 Tapestation. The pool concentration was further determined using Qubit and qPCR methods. Sequencing was performed on an Illumina MiniSeq Mid Output flow cell, and sequencing was initiated using custom sequencing primers targeting the CS1 and CS2 conserved sites in the library linkers. Additionally, the phiX control library was spiked into the run at 15−20% to increase the diversity of the library clustering across the flow cell. After demultiplexing, the percent seen for each sample was used to calculate a new volume to pool for a final sequencing run with evenly balanced indexing across all samples. This pool was sequenced with metrics identical with those of the QC pool. Library preparation and sequencing were performed at the University of Illinois at Chicago Sequencing Core (UICSQC).

NGS Data Preprocessing. For each sample, the NGS reads were first trimmed and filtered using cutadapt (v1.16). Only NGS read pairs with both Illumina Common Sequence adapters, CS1 and CS2, were kept. Of these, CS2 was trimmed off each R1 sequence and CS1 was trimmed off each R2 sequence. Cutadapt parameters were set as follows: a minimum quality cutoff (-q) of 30, a maximum error rate (-e) of 0.05, a minimum overlap (-O) of 10, and a minimum extension length (-m) of 1. The minimum overlap was set to be higher than the default value of 3 because extended sequences in this case are random, and we did not want to filter out sequences where the final 1−10 bases just happen to look like the first 10 bases of CS2 (the read must still contain a full CS2 sequence for it to be kept and subsequently trimmed, however). The 3′ (-a) adapter trimmed from the R1 reads was 5′AGACCAAGTCTCTGCTACCGTA3′ (CS2 reverse complement), and the 5′ (-A) adapter trimmed from the R2 reads was 5′TGTAGAACCATGTCGTCAGTGT3′ (CS1 reverse complement). FastQC was used to quickly inspect the output trimmed fastq files before downstream analysis. See *filter_and_trim_TdT.sh* at https://github.com/tyo-nu/turtles for an example preprocessing script. All runs were trimmed using this script. All initial preprocessing was done on Quest, Northwestern University's high-performance computing facility, using a node running Red Hat Enterprise Linux Server release 7.5 (Maipo) with 4 cores and 4 GB of RAM, although only 1 core was used. Preprocessing took between 5 and 30 min depending on the number of conditions, replicates, and reads per replicate in a given run.

Finally, for each analysis, we did further preprocessing locally. We cut off bases that were still present in the reads but not added during the experiment. Degenerate bases (if any) that are part of the 5′-ssDNA substrate (at its 3′-end before the extension) were removed from the beginning of each sequence. Then, we cut off 5.8 bases off the end of every sequence because we found that, on average, 5.8 bases were being added after the extension reaction during the 16 h ligation step (Figure S9). Because 5.8 is not an integer value, we cut 5 bases off of 80% of the sequences and 6 bases off of 20% of the sequences. We also filtered out sequences with lengths of less than 6 bases.

Time Point Prediction for 0 → 1 Single Step Change Experiment. All further analysis was done in Python using Jupyter Notebooks. You can find all the Jupyter Notebooks used for this publication at https://github.com/tyo-nu/turtles. The following algorithm was applied in order to (1) read and normalize each sequence by its own length, (2) calculate a distance metric using the relative dATP, dCTP, dGTP, and dTTP percent incorporation changes between each condition and the 0 control, and (3) transform distances for all conditions into 0 → 1 space on the basis of the 0 and 1 control distance values.

We first normalized each sequence by length, such that all bases in each sequence were counted across 1000 bins. For example, for a sequence of length 10, the first base would get counted in the first 100 bins, the next base in bins 100−200, and so on.

We then calculated the base composition, X_{ij}, in the sequence for condition i at each bin with position j, using the formula for a closure (eq 1). Note that i is unique for each (condition, replicate) pair if multiple replicates are present for a given experimental condition.

$$X_{ij} = \left[\frac{n_{ijA}}{\sum_{k\in N} n_{ijk}}, \frac{n_{ijC}}{\sum_{k\in N} n_{ijk}}, \frac{n_{ijG}}{\sum_{k\in N} n_{ijk}}, \frac{n_{ijT}}{\sum_{k\in N} n_{ijk}}\right] \tag{1}$$

Here, n_{ijk} is the total count of dATP, dCTP, dGTP, or dTTP depending on the value of k ($k \in N = \{A,C,G,T\}$) across all sequences for condition i at bin j.

To calculate the distance between two compositions at a given bin location (e.g., between the 0 and 1 controls at every bin), we had to first transform the compositional data. We could not simply take the L2 norm difference of each compositional element because the elements of a composition violate the principle of normality due to the total sum rule (all elements add up to 100%). Thus, the data were first transformed by using the center log-ratio (clr) transformation, which maps this 4-component composition from a 3-dimensional space to a 4-dimensional space. We then took the L2 norm of these transformed normal elements. This distance metric is known as the Aitchison distance, which was used here to calculate the base composition distance, $d_j(0,i)$, from the 0 control to each condition i at each bin j (eq 2).

$$d_j(0, i) = \sqrt{\sum_{k\in N}\left[\ln\left(\frac{X_{ijk}}{g(X_{ij})}\right) - \ln\left(\frac{X_{0jk}}{g(X_{0j})}\right)\right]} \tag{2}$$

$N = \{A,C,G,T\}$, and $g(X_{ij})$ is the geometric mean for condition i and bin j across all four bases in N (eq 3).

$$g(X_{ij}) = \sqrt[4]{\prod_{k\in N} X_{ijk}} \tag{3}$$

For condition i and bin j, the output signal s_{ij} is calculated as

$$s_{ij} = \frac{d_j(0, i) - d_j(0, 0)}{d_j(0, 1) - d_j(0, 0)} = \frac{d_j(0, i)}{d_j(0, 1)} \tag{4}$$

where $d_j(0,1)$ is the Aitchison distance between the 0 control base composition and 1 control base composition at bin j. $d_j(0,0) = 0$ for all j. If there were multiple replicates for the 0 control, their average composition was used for X_{0j} (and X_{0jk}) in eq 2. If there were multiple replicates for the 1 control, their average composition was similarly used to calculate $d_j(0,1)$ in eq 4.

Next, the switch times were estimated for each condition i, which contains a change in output signal s_{ij} (e.g., via addition of Co halfway through the reaction). For experiments with more than one change (e.g., 0 → 1 → 0), a more sophisticated approach was used and is detailed below. However, the following simpler, more intuitive approach was used to predict switch times for 0 → 1 and 1 → 0.

Switch times were estimated for a given condition, i, by (1) finding j_i*, the average location across all the sequences (bin position j) at which half the 1 control output signal is reached (i.e., $s_{ij} = 0.5$), (2) calculating α, the ratio of the average rate of nucleotide addition for the 0 and 1 controls, and (3) using j_i* and α to calculate the switch time, t_i*, using eqs 5 and 6. For a derivation of eq 5, see Supplementary Note 3.

https://doi.org/10.1021/jacs.1c07331
J. Am. Chem. Soc. 2021, 143, 16630−16640

$$t_i^* = \frac{\alpha t_{expt}}{\frac{1}{j_i^*} + \alpha - 1} \quad (5)$$

where

$$\alpha = \frac{\overline{r_{a,ctrl}}}{\overline{r_{b,ctrl}}} \quad (6)$$

$\overline{r_{a,ctrl}}$ is the average synthesis rate of the first environmental condition before the switch. For example, $\overline{r_{a,ctrl}}$ would be calculated using the 0 control for the condition 0 → 1, but the 1 control for the condition 1 → 0. The average synthesis rate is calculated by dividing the average extension length by the duration of the experiment. $\overline{r_{b,ctrl}}$ is the average synthesis rate for the second environmental condition (after the switch).

Time Point Estimation for 0 → 1 → 0 Multiple Fluctuation Experiment. To predict the Co^{2+} condition in the 0 → 1 → 0 experiment, we used the algorithm we developed from Glaser et al. for decoding continuous concentrations.[13] The input to this algorithm is the amount of output signal on every nucleotide. Here, the output signal is s_{ij} from the previous section. The algorithm uses this information to predict continuous values of Co^{2+} between 0 and 1 for all time points that are most likely to produce the amount of output signal on the nucleotides. To binarize these predictions, we then set a threshold of 0.5. To be able to predict the values of Co^{2+}, the algorithm requires a knowledge of the expected amount of output signal under the 0 and 1 control conditions. Here, this is the average output signal across nucleotides in the 0 or 1 control experiments. The algorithm also requires a knowledge of the rate of nucleotide addition. Here, we fit an inverse Gaussian distribution to the average experimental dNTP addition rate distribution (the distribution of the sequence lengths divided by the experiment time) from the control experiments. Note that this algorithm also assumes that the rate of dNTP addition is independent of the cation concentration. Thus, when making predictions in the 0 → 1 → 0 experiment, we do not account for differences in the rate of dNTP addition distributions between the 0 and 1 conditions.

***In Silico* Simulations of Recording Faster and Higher Numbers of Input Signal Changes.** Using the average dNTP incorporation rate from experiments, and the amount of output signal in the control conditions, we simulated additional experiments *in silico*. Each simulated experiment had at least 6 signal changes (instances of a single signal change from 0 → 1 or 1 → 0), where each condition was randomly chosen to be 0 or 1. All nucleotides that were added during the 0 or 1 condition had the signal associated with these control conditions. More specifically, to account for the experimental variability in signals within a given control condition, nucleotide signals were sampled from a normal distribution determined by the experimental variability of nucleotide signals within the control conditions. We calculated the variability in two ways, corresponding to the two representative curves in Figure S12A,C. In the first, the variability was calculated across the first 100 nucleotides, in which there were at least 2000 recordings of all base numbers. In the second, the variability was calculated across the first 50 nucleotides, in which there were at least 60000 recordings of all base numbers. Using the output signal of the simulated nucleotides, we used the algorithm we developed from Glaser et al. for decoding binary concentrations.[13] The accuracy corresponds to the percentage of conditions correctly classified as 0 or 1 over the duration of the entire recording experiment.

Time Point Estimation for 0 → 1 and 1 → 0 Single-Step Changes for TURTLES-2 Using an Inverse Model. To predict the Co^{2+} condition in the 0 → 1 → 0 experiment, we used a variation of the algorithm we developed from Glaser et al. for decoding continuous concentrations.[13] This algorithm will predict continuous values of Co^{2+} between 0 and 1 for all time points that are most likely to produce the amount of signal. Here, instead of using the amount of signal on every nucleotide to predict the continuous concentrations, we use the normalized signal.

Let s_{ij} be the signal as a function of the condition i and the normalized position j. Let $\gamma_{ij}(t)$ be the probability that a nucleotide corresponding to the normalized position j was written at time t. Let C_i be the normalized cation concentration for condition i. As in ref 13, our model is that $s_{ij} \approx \sum_t \gamma_{ij}(t) C_i(t)$. We use maximum likelihood estimation to find a C_i value that minimizes $\sum_j (s_{ij} - \sum_t \gamma_{ij}(t) C_i(t))^2$ subject to $C_i(t) \in [0,1]$ and the given condition i. To binarize the predictions, we then set a threshold for C of 0.5.

Here, for an experiment of duration t_{expt} (e.g., 60 min), we let $\gamma_{ij} = N((j - 0.5)/t_{expt}, \sigma)/Z$, where Z renormalizes the probability distribution after values outside the domain of $[0, t_{expt}]$ are set to 0. We set σ for each experiment so that $\gamma_{i1}(t_{expt})$ is equal to the frequency of strands with a single nucleotide divided by t_{expt} (because a normalized position of 1 would generally only be written in the t_{expt}th minute when there is a single nucleotide strand). Note that future work that more accurately models the kinetics of the polymerase to get a more accurate estimate of γ will provide improved results.

When running this algorithm on the Ca^{2+} data which have different rates when Ca^{2+} is or is not present, following the prediction of Ca^{2+} over time with the above algorithm, we used the ratio of incorporation rates between the 0 and 1 conditions, as described by eqs 5 and 6, to rescale the results.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information includes an which are referred to in the main text. The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.1c07331.

Extended descriptions of methods and supplementary figures and tables as described in the text (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Keith E. J. Tyo − *Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, United States;* orcid.org/0000-0002-2342-0687; Phone: +1 847 868 0319; Email: k-tyo@northwestern.edu; Fax: +1 847 491 3728

Authors

Namita Bhan − *Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, United States; Mitolab, Cambridge, Massachusetts 02139, United States*

Alec Callisto − *Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, United States;* orcid.org/0000-0002-9556-7389

Jonathan Strutz − *Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, United States;* orcid.org/0000-0003-0934-4283

Joshua Glaser − *Center for Theoretical Neuroscience, Columbia University, New York, New York 10027, United States*

Reza Kalhor − *Department of Biomedical Engineering, Center for Epigenetics, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, United States;* orcid.org/0000-0002-5558-7545

Edward S. Boyden − *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; McGovern Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Howard Hughes Medical Institute, Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115, United States*

George Church − *Department of Biomedical Engineering, Center for Epigenetics, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, United States*

Konrad Kording − *Department of Neuroscience, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.1c07331

Author Contributions

□N.B. and A.C. contributed equally to this work.

Notes

The authors declare the following competing financial interest(s): A utility patent has been filed for some of the inventions described in this work.

■ ACKNOWLEDGMENTS

The authors acknowledge Marija Milisavljevic for help with some experiments and Bradley Biggs for helpful discussions and comments on the manuscript. This research was supported in part through the computational resources and staff contributions provided for the Quest high-performance computing facility at Northwestern University, which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. All next-generation sequencing was done with the help of the Next Generation Sequencing Core facility at the University of Illinois at Chicago. Sanger sequencing was supported by the Northwestern University NUSeq Core Facility. Gel imaging was supported by the Northwestern University Keck Biophysics Facility and a Cancer Center Support Grant (NCI CA060553). The Keck Biophysics Facility's Azure Sapphire Imager was funded by a 1S10OD026963-01 NIH grant. Protein purification was supported by the Northwestern University Recombinant Protein Production Core. This work was funded by the National Institutes of Health grants R01MH103910 (to K.E.J.T., K.K., E.S.B., and G.C.), and UF1NS107697 (to K.E.J.T., K.K., and E.S.B.) and a National Institutes of Health Training Grant (T32GM008449) through Northwestern University's Biotechnology Training Program (to J.S. and A.C.). NGS data are available from Sequence Read Archive: https://www.ncbi.nlm.nih.gov/sra/PRJNA542184.

■ REFERENCES

(1) Church, G. M.; Gao, Y.; Kosuri, S. Next-Generation Digital Information Storage in DNA. *Science (Washington, DC, U. S.)* **2012**, *337*, 1628−1628.

(2) Goldman, N.; et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **2013**, *494*, 77−80.

(3) Erlich, Y.; Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science (Washington, DC, U. S.)* **2017**, *355*, 950−954.

(4) 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.* **2015**, *54*, 2552−2555.

(5) Sheth, R. U.; Yim, S. S.; Wu, F. L.; Wang, H. H. Multiplex recording of cellular events over time on CRISPR biological tape. *Science* **2017**, *358*, 1457−1461.

(6) Loveless, T. B.; et al. Lineage tracing and analog recording in mammalian cells by single-site DNA writing. *Nat. Chem. Biol.* **2021**, *17*, 739−747.

(7) Shipman, S. L.; Nivala, J.; Macklis, J. D.; Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science (Washington, DC, U. S.)* **2016**, *353*, aaf1175.

(8) 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* **2017**, *547*, 345−349.

(9) Palluk, S.; et al. De novo DNA synthesis using polymerase-nucleotide conjugates. *Nat. Biotechnol.* **2018**, *36*, 645−650.

(10) Lee, H. H.; Kalhor, R.; Goela, N.; Bolot, J.; Church, G. M. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nat. Commun.* **2019**, *10*, 2383.

(11) Lee, H.; et al. Photon-directed multiplexed enzymatic DNA synthesis for molecular digital data storage. *Nat. Commun.* **2020**, *11*, 1−9.

(12) Yim, S. S.; et al. Robust direct digital-to-biological data storage in living cells. *Nat. Chem. Biol.* **2021**, *17*, 246−253.

(13) Glaser, J. I. Statistical Analysis of Molecular Signal Recording. *PLoS Comput. Biol.* **2013**, *9*, e1003145.

(14) Kording, K. P. Of toasters and molecular ticker tapes. *PLoS Comput. Biol.* **2011**, *7*, e1002291.

(15) Kelman, Z.; O'Donnell, M. DNA polymerase III holoenzyme: Structure and function of a chromosomal replicating machine. *Annu. Rev. Biochem.* **1995**, *64*, 171−200.

(16) Motea, E. A.; Berdis, A. J. Terminal Deoxynucleotidyl Transferase: The Story of a Misguided DNA Polymerase. *Biochim. Biophys. Acta* **2015**, *21*, 253−260.

(17) Boulé, J. B.; Johnson, E.; Rougeon, F.; Papanicolaou, C. High-level expression of murine terminal deoxynucleotidyl transferase in Escherichia coli grown at low temperature and overexpressing argU tRNA. *Mol. Biotechnol.* **1998**, *10*, 199−208.

(18) Chang, L M; Bollum, F J Multiple Roles of Divalent Deoxynucleotidyltransferase Cation in the Terminal Reaction *. *J. Biol. Chem.* **1990**, *265*, 17436−17440.

(19) Fowler, J. D.; Suo, Z. Biochemical, Structural, and Physiological Characterization of Terminal Deoxynucleotidyl Transferase. *Chem. Rev.* **2006**, *106*, 2092−2110.

(20) Deibel, M. R.; Coleman, M. S. Biochemical properties of purified human terminal deoxynucleotidyltransferase. *J. Biol. Chem.* **1980**, *255*, 4206−12.

(21) Grienberger, C.; Konnerth, A. Imaging Calcium in Neurons. *Neuron* **2012**, *73*, 862−885.

(22) Whitaker, M. Calcium at fertilization and in early development. *Physiol. Rev.* **2006**, *86*, 25−88.

(23) Stricker, S. A. Comparative Biology of Calcium Signaling during Fertilization and Egg Activation in Animals. *Dev. Biol.* **1999**, *211*, 157−176.

(24) Rosenberg, S. S.; Spitzer, N. C. Calcium Signaling in Neuronal Development. *Cold Spring Harbor Perspect. Biol.* **2011**, *3*, a004259.

(25) Frederickson, C. J.; Koh, J.-Y.; Bush, A. I. The neurobiology of zinc in health and disease. *Nat. Rev. Neurosci.* **2005**, *6*, 449−462.

(26) Motea, E. A.; Berdis, A. J. Terminal deoxynucleotidyl transferase: The story of a misguided DNA polymerase. *Biochim. Biophys. Acta, Proteins Proteomics* **2010**, *1804*, 1151−1166.

(27) Nakai, J.; Ohkura, M.; Imoto, K. A high signal-to-noise ca2+ probe composed of a single green fluorescent protein. *Nat. Biotechnol.* **2001**, *19*, 137−141.

(28) Palmer, A. E.; et al. Ca2+Indicators Based on Computationally Redesigned Calmodulin-Peptide Pairs. *Chem. Biol.* **2006**, *13*, 521−530.

(29) Edwardraja, S. Caged activators of artificial allosteric protein biosensors. *ACS Synth. Biol.* **2020**, *9*, 1306.

(30) Guo, Z.; et al. Generalizable Protein Biosensors Based on Synthetic Switch Modules. *J. Am. Chem. Soc.* **2019**, *141*, 8128.

(31) Smith, M. A.; Arnold, F. H. Designing libraries of chimeric proteins using SCHEMA recombination and RASPP. *Methods Mol. Biol.* **2014**, *1179*, 335−343.

(32) Crotti, L.; et al. Calmodulin mutations associated with recurrent cardiac arrest in infants. *Circulation* **2013**, *127*, 1009−1017.

(33) Zulkifli, S. N.; Rahim, H. A.; Lau, W. J. Detection of contaminants in water supply: A review on state-of-the-art monitoring technologies and their applications. *Sens. Actuators, B* **2018**, *255*, 2657−2689.
(34) Slomovic, S.; Pardee, K.; Collins, J. J. Synthetic biology devices for in vitro and in vivo diagnostics. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 14429−14435.
(35) Liu, X.; et al. Design of a transcriptional biosensor for the portable, on-demand detection of cyanuric acid. *ACS Synth. Biol.* **2020**, *9*, 84−94.
(36) Jung, J. K.; et al. Cell-free biosensors for rapid detection of water contaminants HHS Public Access Author manuscript. *Nat. Biotechnol.* **2020**, *38*, 1451−1459.
(37) Thavarajah, W.; et al. Point-of-Use Detection of Environmental Fluoride via a Cell-Free Riboswitch-Based Biosensor. *ACS Synth. Biol.* **2020**, *9*, 10−18.
(38) Marblestone, A. H.; et al. Physical principles for scalable neural recording. *Front. Comput. Neurosci.* **2013**, *7*, 1−34.
(39) Marblestone, A. H.; Daugharthy, E. R.; Kalhor, R.; Peikon, I. D.; Kebschull, J. M.; Shipman, S. L.; Mishchenko, Y.; Lee, J. H.; Kording, K. P.; Boyden, E. S.; Zador, A. M.; Church, G. M. Rosetta Brains: A Strategy for Molecularly-Annotated Connectomics. *ArXiv: Neurons and Cognition* **2014**, 1−18.
(40) Vikesland, P. J. Nanosensors for water quality monitoring. *Nat. Nanotechnol.* **2018**, *13*, 651−660.
(41) Long, F.; Zhu, A.; Shi, H.; Wang, H.; Liu, J. Rapid on-site/in-situ detection of heavy metal ions in environmental water using a structure-switching DNA optical biosensor. *Sci. Rep.* **2013**, *3*, 2308.
(42) Eisenstein, M. Enzymatic DNA synthesis enters new phase. *Nat. Biotechnol.* **2020**, *38*, 1113−1115.
(43) Akerboom, J.; et al. Optimization of a GCaMP Calcium Indicator for Neural Activity Imaging. *J. Neurosci.* **2012**, *32*, 13819−13840.
(44) Chen, T.-W.; et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **2013**, *499*, 295−300.
(45) Green, M. R.; Sambrook, J. Precipitation of DNA with Ethanol. *Cold Spring Harb. Protoc.* **2016**, *2016*, pdb.prot093377.
(46) de Paz, A. M.; et al. High-resolution mapping of DNA polymerase fidelity using nucleotide imbalances and next-generation sequencing. *Nucleic Acids Res.* **2018**, *46*, e78−e78.
(47) Zamft, B. M. Measuring cation dependent DNA polymerase fidelity landscapes by deep sequencing. *PLoS One* **2012**, *7*, e43876.

https://doi.org/10.1021/jacs.1c07331
J. Am. Chem. Soc. 2021, 143, 16630−16640