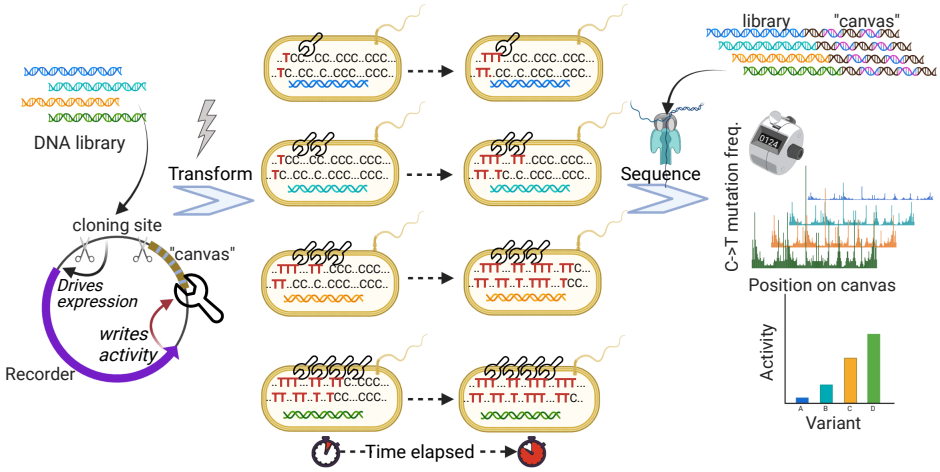
# High-throughput molecular recording can determine the identity and biological activity of sequences within single cells

Boqiang Tu[1,2] and Kevin M. Esvelt[2]

[1]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

**Abstract**

Methods of measuring the activities of biomolecules and pathways are fundamental to the study and engineering of biological systems. Traditional marker genes that generate light or color are limited in their dynamic range and often require expensive equipment. Here we describe a way to simultaneously detect the identity and measure the activity of each molecular variant within a mixed population of cells using nanopore sequencing. By linking the desired activity to expression of a base editor that induces mutations in a "canvas" adjacent to the DNA encoding the molecule or pathway of interest, we show that a single long-read sequencing step can retrieve identity and activity at numerous timepoints in bacteria, potentially greatly increasing the available dynamic range. Our technique can replace flow cytometry in directed evolution and may be capable of directly mapping biomolecular fitness landscapes.



# Introduction

The field of synthetic biology advances by generating novel molecular tools and using them to build increasingly complex and sensitive gene regulatory networks. Both tool development and network characterization require methods of experimentally measuring the activity of each candidate design. Unfortunately, the reproducibility, dynamic range, and throughput of current measurement methods are limited [1] by the signal that can be generated by the continuous production of a fluorescent, luminescent, or colorigenic protein. While flow cytometry can measure the activity of single cells, it requires either sequential loading of samples onto the flow cytometer, which severely limits the number of biological parts screened in a single experiment, or sequencing a library before and after sorting into a series of bins by activity level and inferring real activities by evaluating enrichment and the relative frequency of library members across the bins.

Here we present Direct High-throughput Activity Recording and Measurement Assay (DHARMA), a method of precisely measuring the activity of each member of a genetically encoded library over a wide dynamic range using only nanopore sequencing. Unlike previous molecular recording methods that record external stimuli at the population level, DHARMA can measure the activity of each individual cell within a mixed population. Our implementation utilizes a base editor as a reporter protein to record the activity of a functional sequence onto a DNA "canvas" contiguous to each library member. The information recorded in the canvas and the identity of the contiguous library member are then subjected to nanopore sequencing. After demultiplexing, each input sequence is assigned an activity score based on the normalized number of C to T mutations accumulated in the canvas.

To test this concept and begin calibrating our method, we initially designed plasmid constructs consisting of a base editor expressed under the control of a constitutive promoter, a canvas with multiple target sites for activity recording, and a GFP or luxAB expressed under the control of the same promoter to serve as a signal reference for relative activity quantification. In the presence of sgRNA, such constructs simultaneously measure the promoter strength using both sequencing and either fluorescence or luminescence intensity. To challenge the system with a wide range of activity levels, we cloned 24 previously characterized promoters and showed that nanopore sequencing alone is sufficient for activity quantification and in principle offers a greater dynamic range than is achievable with fluorescence or luminescence.

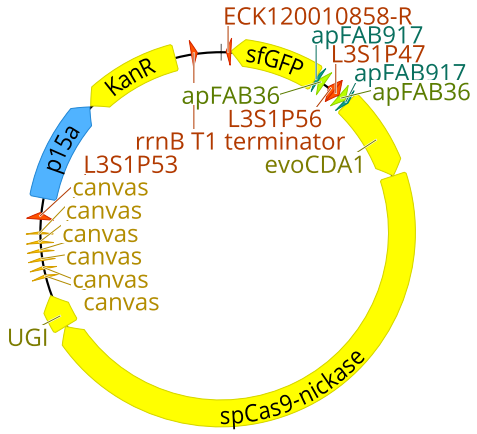
# Results

## Design, construction and initial characterization of a molecular recording system for activity reporting

Base editors are a class of molecular machinery that introduces single-base transitions to a specifically targeted DNA sequence without causing double strand breaks[2] that have previously been used as population-level molecular recorders to detect external stimuli[3]. We hypothesized that a similar molecular recorder could function at the single-cell level to measure the activity of genetically encoded signals. As long as the concentration of guide RNA is not limiting, the expression level of a base editor within each individual cell should determine the frequency of mutations in a targeted repetitive region, which we term the "canvas", which should accumulate over time. Modeling the relationship between base editor expression level and the measured mutation profile could be used to quantify the absolute activity of any gene-encoded sequence whose activity can be coupled to production of the base editor. Sequencing samples exposed to the base editor for a brief window could differentiate between highly active and active sequences before saturation of the canvas, while sequencing samples exposed for many hours or days could differentiate between marginally and negligibly active sequences.

To test and validate a system in which the molecular activity of a functional sequence is recorded in a way such that the activity and the identity of the functional sequence can be obtained concomitantly via long-read sequencing, we designed a plasmid construct (Figure 1a) consisting of a highly active cytosine base editor[4], a functional sequence of interest and a canvas, which we refer to as the recorder plasmid (RP). We chose the evoCDA1 cytosine base editor due to its apparent distinction as the most highly active method of generating targeted mutations[5] and further improved its sensitivity by introducing multiple target sites in the canvas. While in principle other mutation-generating systems could also be used, the high activity and multi-site targeting of base editing is particularly suitable. For the purpose of system validation, we

(a)




(b)

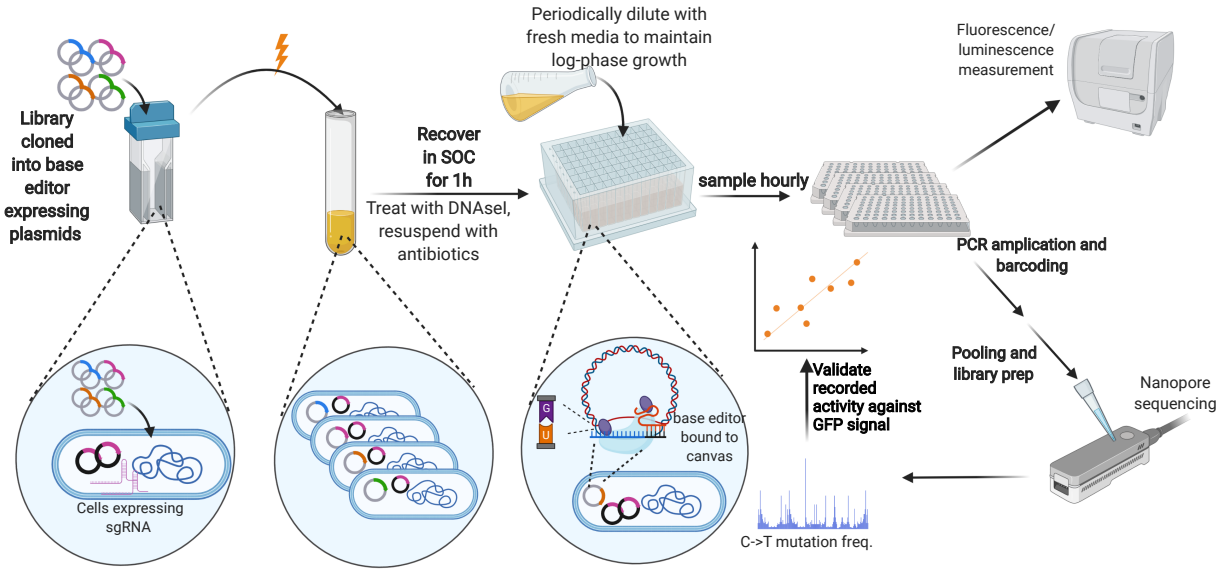



Figure 1: Design of plasmid constructs and general workflow. (a) Annotation of a typical "recorder plasmid (RP)" construct used for system validation. Base editor and GFP are transcribed in opposite directions and insulated with a strong terminator upstream in each direction. Identical promoters and RBSs used for both coding sequences allow calibration of base editor activity in the form of mutations recorded against fluorescence signal corresponding to such activity. (b) General workflow of DHARMA.

chose the promoter driving the expression of the base editor as our functional sequence of interest, which was insulated from potential upstream read-through using a strong synthetic terminator. Upstream of the base editor cassette and in the reverse direction, an identical promoter was used to drive the transcription of GFP. Identical ribosome binding site (RBS) sequences were used for both coding sequences such that the activity of the promoter can be independently measured via fluorescence as a source of validation, an arrangement that proved more reliable than attempting to encode a fluorescent gene and the base editor in series (data not shown). Promoter and RBS sequences were synthetic and extensively characterized in previous works[6]. Our canvas sequence comprises either 5 or 10 repeats of an identical cytosine-rich CRISPR-Cas9 target, selected from a list of optimized targets with maximum editing efficiency [7]. For high-throughput testing of a library of functional sequences, the region flanked by RBS apFAB917 was replaced with a lacZ$\alpha$ cassette flanked by two PaqCI restriction sites to allow scarless insertion of library members.

To minimize the impact of varying timing and quantity of sgRNA expression on the accumulation of mutations in the canvas, we used a separate plasmid to constitutively express sgRNA in cells that were later made electrocompetent. To probe the possible confounding effect of variation in the amount of sgRNA on mutation frequency on the canvas, we generated cells expressing sgRNA driven by either a strong or a weak promoter and introduced RPs into these cells via electroporation. After cells were allowed to recover for 1h, we immediately started liquid cultures with appropriate antibiotics. This allowed us to sample the cultures and take snapshots of the base editing activities in the first few hours before the number of C to T mutations on some canvas reached saturation. To enable multiplexed sequencing of multiple sample collected at different time points, we amplified the region of interest with barcoded primers directly from the liquid culture and successfully sequenced up to 480 samples on a single MinION R10.4 flow cells to yield sufficient number reads for quantification of mutations on canvas. Interestingly, the strength of the promoter driving the expression of sgRNA did not have a pronounced impact on the level of base editing activity regardless of the activity of the promoter driving the expression of the base editor, which might indicate that the sgRNA produced was in excess, and the concentration of base editor plays the predominant role in determining the number of mutations in the canvas region. As shown in Figure S1, C to T mutations accumulated rapidly in the first 8 hours after electroporation and at a higher rate in cells where a strong promoter drives the expression of the base editor and the GFP gene, the fluorescence intensity of which confirms the relative strengths of promoters. The rate of increase in number of mutations slowed in the next 10 hrs as the number of cytosines available for deamination decreased. Note that the cells were allowed to grow to stationary phase overnight, which might have slowed metabolic activities in general.

## Nanopore sequencing workflow and computational pipeline

As the activity is recorded in a region contiguous to the functional sequence, it is critical to obtain full-length reads that contain both sequences such that a direct mapping between activity and sequence identity can be established. Nanopore sequencing technology has now been frequently used to generate full-length reads>20kb and the latest chemistry yields a modal raw read accuracy of 99%. Full-length reads of RPs or amplicons of RPs obtained from nanopore sequencing are sufficiently accurate for mutation frequency quantification but present challenge in identifying point mutations or other minor variations within the functional sequence of interest. To this end, we included a 24-base barcode adjacent the sequence of interest during synthesis. These barcodes were used in previous works to demultiplex up to 96 samples with 95% accuracy [8]. For the purpose of system validation, we also performed barcoding PCR directly over the canvas region as an alternative multiplexing method, which was possible in this case because the identity of the functional sequence was known. In the scenario in which the identities of functional sequences cannot be readily associated with an additional identifier, reads can be first grouped by exact matches in the functional sequence region and then further clustered based on error tolerance set for a particular application. These clustered reads are treated as demultiplexed read groups each containing reads with identical functional sequences. Within each read group, the canvas region of each read was aligned to the reference sequence using the Smith–Waterman algorithm [9], and the number of C to T mutations was counted at each position within the canvas region and normalized to the total number of full-length reads in the read group. We generated a mutation profile for each read group and computed the area under the curve as the primary metric of base editing activity. While we observed some degree of noise, potentially caused by sequencing error and non-specific deamination, it was unlikely to materially impact the activity readout due to its low

amplitude. These results validated the basic method.

## Validation of molecular recording-based activity measurement using a library of 24 promoters

To demonstrate that this molecular recording-based approach can accurately quantify activity over a wide dynamic range, we selected 24 synthetic promoters (Table S1)), the relative strengths[6] of which were previously determined to vary up to 100-fold. An 1000-fold dynamic range in protein expression can be achieved in combination with RBS with varying activities. To minimize the confounding effects of adjacent sequences, we included 8 insulated promoters among the 24 promoters chosen for system validation[10]. For the purpose of system validation, we separately cloned each promoter into a backbone comprising GFP, base editor, and canvas (Figure 1a), and individually transformed each of the resulting constructs into cells expressing sgRNA driven by a strong constitutive promoter via electroporation. As strong base editing activity can lead to saturation of the canvas in overnight cultures (data not shown), we reasoned that bypassing the conventional post-transformation colony selection process and establishing liquid cultures immediately after post-electroporation outgrowth would allow us to capture the base editing events at early time points. To prevent the loss of plasmids, we added appropriate antibiotics directly to liquid cultures after one hour of outgrowth. Such cultures were maintained at constant density under log-phase growth condition using an automated liquid handler and sampled hourly for GFP assay and nanopore sequencing.

The relative strengths of promoters measured by GFP fluorescence intensity were largely consistent with those described in previous reports (data not shown). During a 42-hour period of incubation after electroporation, we observed an accumulation of C to T transitions in the canvas region over time among the majority of samples, with faster accumulation in samples with stronger promoters driving expression of the base editor. Among all the samples collected at different time points, the normalized C to T mutation frequency exhibited a wide numerical range from 500 to over 50000 mutations/1000 reads, and the mutation frequency appeared to be an increasing function of promoter activity represented by GFP fluorescence intensity (Figure 2a). Data smoothing of the time series was performed by taking the moving average with one neighboring data point. These time series of normalized mutation frequencies were then fitted to a generalized logistic model to assess the kinetics of base editor-mediated activity recording. Linear regression showed a strong correlation between the maximum rate of mutagenesis and the log-transformed GFP intensity ($R^2 = 0.96$) among insulated promoters (Figure 2d)). These eight promoters (Figure 2c) also produced strong correlation between log-transformed mutation frequency and log-transformed GFP fluorescence intensity ($R^2 > 0.9$) at several different time points.

To model the numerical relationship between GFP signal and mutation frequency for all samples, we plotted the log-transformed mutation frequency observed in each promoter condition against the log-transformed normalized GFP fluorescence intensity measured at steady state in saturated cultures near the time points where $V_{\max}$ was reached in most samples. As shown in Figure S2b, there was an apparent linear relationship between these two variables ($R^2 = 0.79$) at 8h after electroporation. At early time points, low-activity promoters produced indistinguishable numbers of mutations on canvas while high-activity promoters resulted mutations as apparent increasing functions of promoter strengths. At later time points, the relationship between mutation frequency and promoter strength became observable among low-activity promoters while high-activity promoters led to saturation of C to T mutations in the canvas region (Figures 2a,S2). These data suggest that the quantitative relationship between the mutation frequency in the canvas region and the underlying activity of functional sequence can be described with a tractable model over a dynamic range of at least 100-fold difference. Given the cumulative nature of base editing on canvas, a much higher dynamic range is likely to achieved if, cells with low-activity functional sequences are allowed to grow in continuous culture for extended period of time until the mutation frequency reaches a range of reliable quantification, a process similar to gain adjustment in a wide range of measurement techniques. Otherwise, the number of sgRNA target repeats in the canvas can be increased and baseline expression level of base editor can be tuned to increase the sensitivity of the activity recording mechanism.

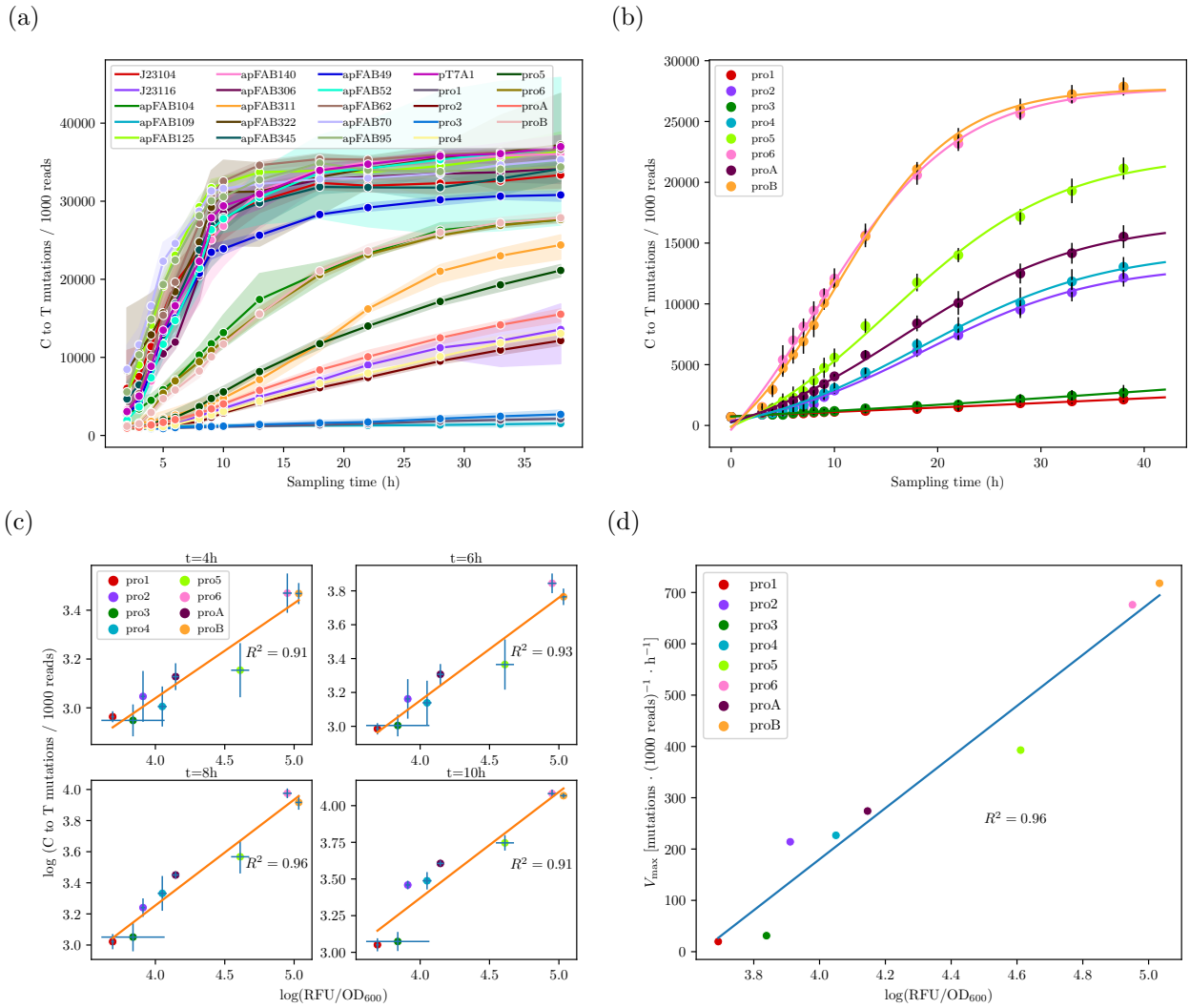



Figure 2: Validation of DHARMA performance across 24 promoters with a wide range of activities. DNA fragments containing 24 promoters were synthesized and cloned into a plasmid as shown in Figure 1, such that identical promoters drove the expression of base editor and GFP. The resulting constructs were separately transformed into electrocompetent NEB 10-beta cells which already harbor a sgRNA expression plasmid. After 1 h of outgrowth, cells were treated with DNaseI and resuspended in media containing chloramphenicol and kanamycin. Samples were taken approximately every hour after outgrowth for fluorescence measurement and barcoding PCR amplification of the canvas sequences. (a) Accumulation of C to T mutations in the canvas region of each construct with a different promoter driving expression of base editor. Canvas regions were amplified using barcoded primers and sequenced on a MinION R10.4 flow cell. Demultiplexed reads were aligned to the reference canvas sequence and numbers of C to T mismatches were counted and normalized against total number of full-length reads for each sample. Variable volumes of culture were replaced with fresh media every 20-30 min on an automated liquid handling platform to maintain log-phase growth. (b) As in (a), with data from insulated promoters fitted to generalized logistic functions. Mutation rate was set to 500 at $t = 0$ to account for baseline mutations and sequencing errors, (c) The log-transformed total number of C to T mutations in the canvas region was plotted against the corresponding log-transformed GFP fluorescence intensity for each insulated promoter at different sampling time. (d) $V_{\max}$ calculated from fitted curves in (b) plotted against log-transformed fluorescence intensity. Error bars represent standard deviation.

# Discussion

We have developed Direct High-throughput Activity Recording and Measurement Assay (DHARMA), a sequencing-based system for high-throughput characterization of DNA-encoded biomolecular activities with many advantages over existing mainstream methods such as flow cytometry. Based on the principle of molecular recording, DHARMA writes a log of activity history onto a segment of DNA ("canvas") contiguous to the activity-encoding sequence, and this information is later read via nanopore sequencing and decoded using a bioinformatics algorithm. The physical linkage between the activity-encoding sequenceand canvas creates a reliable mapping between the identity of the functional sequence and the associated activity in a specific biochemical context, enabling highly multiplexed measurement of DNA-encoded activities in a single nanopore sequencing run. A large number of diverse DNA sequences can be pooled and directly used input for activity measurement without the need of sequential loading of samples as required for accurate and precise measurement using methods such as FACS. Furthermore, given sufficient read depth, DHARMA provides definitive activity information of every individual variant in the library as opposed to multiple distributions ("bins") of activities resulting from a group of functionally similar variants, which require subsequent sequencing for identification. Compared to "binning", measurement on single-variant level is a superior representation of true activity, in particular when the dynamic range of activities is high and the number of "bins" are physically limited.

Many laboratories have contributed to the thriving field of DNA-based memory devices, which have been widely used to record external stimuli [11, 12, 13, 14, 15, 16] or cell lineage information [17, 18, 19] in the genomes of cell populations using a variety of mutagenic methods. High-precision analog signal recording has recently been achieved by CRISPR-mediated analog multi-event recording apparatus (CAMERA)[3], which utilizes base editors to enable multiplexed recording of cellular states in response to environmental stimuli. While these devices demonstrated the feasibility of event recording using DNA as a storage medium in a population of cells, they did not explore the application of DNA-based recording in profiling the molecular activities of individual library members. DHARMA enables highly multiplexed measurements of biological activities at the single-variant level within a mixed population of cells. Compared to conventional techniques such as FACS, DHARMA improves throughput by orders of magnitude and achieves high precision without laborious calibration or expensive equipment. With further validation and development, we envision broad applications of this system in the fields of directed evolution, synthetic biology, protein engineering, epigenetics and biological drug discovery.

# Methods

## Plasmid construction and growth conditions

Multilevel Golden Gate cloning and Gibson cloning were used to construct all plasmids used in the experiments. Plasmids were assembled from modules each containing a scarless transcription unit insulated by strong terminators. T4 DNA ligase and Type IIS endonucleases, BsaI, BsmBI, and PaqCI (New England Biolabs, Ipswitch, MA) were used in different levels of modular assembly. To generate a library of plasmids each containing a different promoter, a lacZ$\alpha$ cassette containing PaqCI sites was inserted between GFP and base editor RBS sequences via restriction digestion and Gibson cloning. Synthetic dsDNA fragments (Integrated DNA Technologies, Coralville, IA) each containing a different promoter and/or a bidirectional terminator, and a 24-base barcode were inserted scarlessly to replace the lacZ$\alpha$ cassette. Blue-white screening was performed as per manufacturer's instructions. NEB 10-beta cells (New England Biolabs, Ipswitch, MA) were used for cloning and testing of all constructs. Transformation and selection conditions were based on manufacturer's recommendations. After plasmids were successfully cloned, all cells were grown in Davis Rich Media [20] to lower background noise in fluorescence measurements.

## Activity recording and sample collection

Commercial competent cells were transformed with plasmids carrying a canvas repeat sequence-targeting sgRNA cassette driven by strong constitutive promoter apFAB36 [21], and rendered electrocompetent using standard procedures. Recorder plasmids were introduced either individually or as a library into the cells

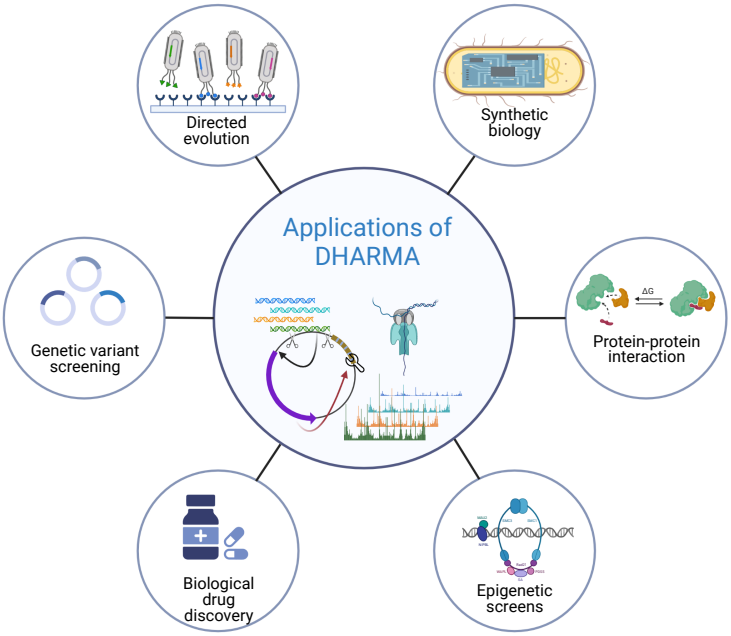



Figure 3: Future applications of DHARMA

above via electroporation at 1700V. Cells were immediately resuspended in SOC medium and allowed to recover at 37°C for 1h. To eliminate plasmids that did not migrate into the cells during electroporation, cells were pelleted, washed with DNaseI reaction buffer (New England Biolabs, Ipswitch, MA), resuspended in DNaseI buffer, and incubated at 37°C for 10min with 2U of DNaseI. Cells were then resuspended in DRM supplemented with appropriate antibiotics at a density of approximately 0.5 $OD_{600}$, and maintained at this density at 37 °C by periodically diluting the cultures with fresh DRM supplemented with appropriate antibiotics. One hundred microliters of cultures were collected hourly for GFP fluorescence measurement and PCR amplification for downstream sequencing. Samples were immediately cooled to 4°C to stop base editing activities.

## Reporter assay and Nanopore sequencing

Samples collected as described in the last section were washed with 10mM HEPES buffer and loaded onto a plate reader (BMG Labtech, Ortenberg, Germany) for fluorescence measurement at 470/515 nm (Ex/Em). Absorbance at 600nm was also measured for fluorescence normalization across different cell densities. To amplify the canvas and activity-encoding regions on the recorder plasmid construct,0 5µL of culture sample was directly used in a 10µL PCR reaction using PrimeStar Max master mix (Takara Bio, San Jose, CA) under conditions recommended by the manufacturer. Primers (Azenta Life Sciences, Chelmsford, MA) used in the reactions include 24-base barcodes on the 5′ end to allow highly multiplexed sequencing across different time points on a single flow cell. PCR reactions were pooled and purified with magnetic beads (Aline Biosciences, Woburn, MA) at bead suspension to sample ratio of 0.8-1x. Sequencing library was prepared using the SQK-LSK112 Ligation Sequencing Kit and sequenced on a R10.4 flow cell as per manufacturer's instructions (Oxford Nanopore Technologies, Oxford, UK).

## Data Analysis

Basecalling and demultiplexing were performed using Guppy v5.0.7 (Oxford Nanopore Technologies, Oxford, UK) with the high-accuracy model shipped with the software. Consensus calling was performed within each group of demultiplexed reads using pbdagcon (`https://github.com/PacificBiosciences/pbdagcon`) to identify or confirm the activity-encoding sequence. Demultiplexed raw reads were then truncated and aligned

to the reference sequence of the repetitive canvas region using the Smith-Waterman algorithm implemented in Julia [22]. For each read, the occurrences of mismatch where cytosine was replaced by thymidine and their positions within the canvas region were stored in a binary vector, the index of which corresponds to position relative to the first base of canvas. The arithmetic sum of these vectors within each demultiplexed read group was normalized with the total number of full-length reads with the group to yield the mutation profile of the canvas sequences associated with a particular library member or variant. The area under curve (AUC) or the total number of mutations was computed to yield the metric for single-variant level activities. Samples with apparent demultiplexing or amplification problems were excluded. Data smoothing was performed by taking moving average with one neighboring data point. Time series mutation rate data were fitted to a generalized logistic function using the Levenberg–Marquardt algorithm. To validate system performance against independently measured activities, log-transformed AUC for each variant was plotted against corresponding log-transformed fluorescence intensity. Linear regression was performed on the log-transformed data using the least-squares algorithm.

Figures 1a and S3 were created using Geneious Prime (`http://www.geneious.com`). Graphical abstract and Figure 1b were created using Biorender (`http://www.biorender.com`).

## Acknowledgments

The authors would like to thank Brian Wang and Emma Chory for their constructive and generous feedback on this project. Special thanks are due to Summer DeAmelio for her extraordinary technical assistance on nanopore sequencing. We are deeply grateful for support from the Open Philanthropy Project, Reid Hoffman, and the MIT Media Lab. None of the funders played any role in the design, execution, analysis, writing, or decision to submit results.

## Competing interests

B.T. and K.M.E. are the authors of a provisional patent filed by MIT on the method.

## Supplementary materials

| Promoter | Sequence |
|---|---|
| pro1 | GGCGCACAGCTAACACCACGTCGTCCCTATCTGCTGCCCTAGGTCTATGA<br>GTGGTTGCTGGATAACTTTACGGGCATGCATAAGGCTCGGTATCTATAT<br>TCAGGGAGACAACAACGGTTTCCCTCTACAAATAATTTTGTTTAACTTT |
| pro2 | GGCGCACAGCTAACACCACGTCGTCCCTATCTGCTGCCCTAGGTCTATGA<br>GTGGTTGCTGGATAACGCGGTGGGCATGCATAAGGCTCGTAGGCTATAT<br>TCAGGGAGACAACAACGGTTTCCCTCTACAAATAATTTTGTTTAACTTT |
| pro3 | GGCGCACAGCTAACACCACGTCGTCCCTATCTGCTGCCCTAGGTCTATGA<br>GTGGTTGCTGGATAACTTTACGGGCATGCATAAGGCTCGGAGGATATAT<br>TCAGGGAGACAACAACGGTTTCCCTCTACAAATAATTTTGTTTAACTTT |
| pro4 | GGCGCACAGCTAACACCACGTCGTCCCTATCTGCTGCCCTAGGTCTATGA<br>GTGGTTGCTGGATAACTTTACGGGCATGCATAAGGCTCGGATGATATAT<br>TCAGGGAGACAACAACGGTTTCCCTCTACAAATAATTTTGTTTAACTTT |
| pro5 | GGCGCACAGCTAACACCACGTCGTCCCTATCTGCTGCCCTAGGTCTATGA<br>GTGGTTGCTGGATAACTTTACGGGCATGCATAAGGCTCGTAGGATATAT<br>TCAGGGAGACAACAACGGTTTCCCTCTACAAATAATTTTGTTTAACTTT |
| pro6 | GGCGCACAGCTAACACCACGTCGTCCCTATCTGCTGCCCTAGGTCTATGA<br>GTGGTTGCTGGATAACTTTACGGGCATGCATAAGGCTCGTAAAATATAT<br>TCAGGGAGACAACAACGGTTTCCCTCTACAAATAATTTTGTTTAACTTT |

| | |
|---|---|
| proA | GGCGCACAGCTAACACCACGTCGTCCCTATCTGCTGCCCTAGGTCTATGA GTGGTTGCTGGATAACTTTACGGGCATGCATAAGGCTCGTAGGCTATAT TCAGGGAGACAACAACGGTTTCCCTCTACAAATAATTTTGTTTAACTTT |
| proB | GGCGCACAGCTAACACCACGTCGTCCCTATCTGCTGCCCTAGGTCTATGA GTGGTTGCTGGATAACTTTACGGGCATGCATAAGGCTCGTAATATATAT TCAGGGAGACAACAACGGTTTCCCTCTACAAATAATTTTGTTTAACTTT |
| apFAB49 | GGCGCGCCAAAAAGAGTATTGACTTTTATCCCTTGCGGCGAATACTTACA GCCATGTAGA |
| apFAB52 | GGCGCGCCAAAAAGAGTATTGACTTTTATCCCTTGCGGCGACATAATTAT TTCATAGTTC |
| apFAB62 | GGCGCGCCTTGACAATTAATCATCCGGCTCGCATAATTATTTCATTTCAG |
| apFAB70 | GGCGCGCCTTGACATCGCATCTTTTTGTACCTATAATGTGTGGATAGAGT |
| apFAB95 | GGCGCGCCAAAAAATTTATTTGCTTTCGCATCTTTTTGTACCTATAATGT GTGGATAATAA |
| apFAB104 | GGCGCGCCTCGACATAAAGTCTAACCTATAGGATACTTACAGCCATAGCTT |
| apFAB109 | GGCGCGCCTCGACAATTAATCATCCGGCTCGATACTTACAGCCATCGATT |
| apFAB125 | GGCGCGCCTCGACATTTATCCCTTGCGGCGATATAATGTGTGGATAATCC |
| apFAB140 | GGCGCGCCCACGGTGTTAGACATCAGGAAAATTTTTCTGTATAATGTGTG GATGCTTA |
| apFAB306 | GGCGCGCCTTGACAATTAATCATCCGGCTCGTAGTGTTTGTGGATGTTG |
| apFAB311 | GGCGCGCCTTGACAATTAATCATCCGGCTCGTAGGTTGTGTGGACGGCT |
| apFAB322 | GGCGCGCCTTGCGTATTAATCATCCGGCTCGTATAATGTGTGGATGATC |
| apFAB345 | GGCGCGCCTTGACAATTAATCATCCGGCTCGTAGAGTATGTGGAGTATC |
| BBa_J23104 | GGCGCGCCTTGACAGCTAGCTCAGTCCTAGGTATTGTGCTAGCTTACG |
| BBa_J23116 | GGCGCGCCTTGACAGCTAGCTCAGTCCTAGGGACTATGCTAGCAGGAT |
| pT7A1 | GGCGCGCCTCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTT ACAGCCATCGAGAGCTGCG |

Table S1: Promoters used in this paper for recorder plasmid construct validation

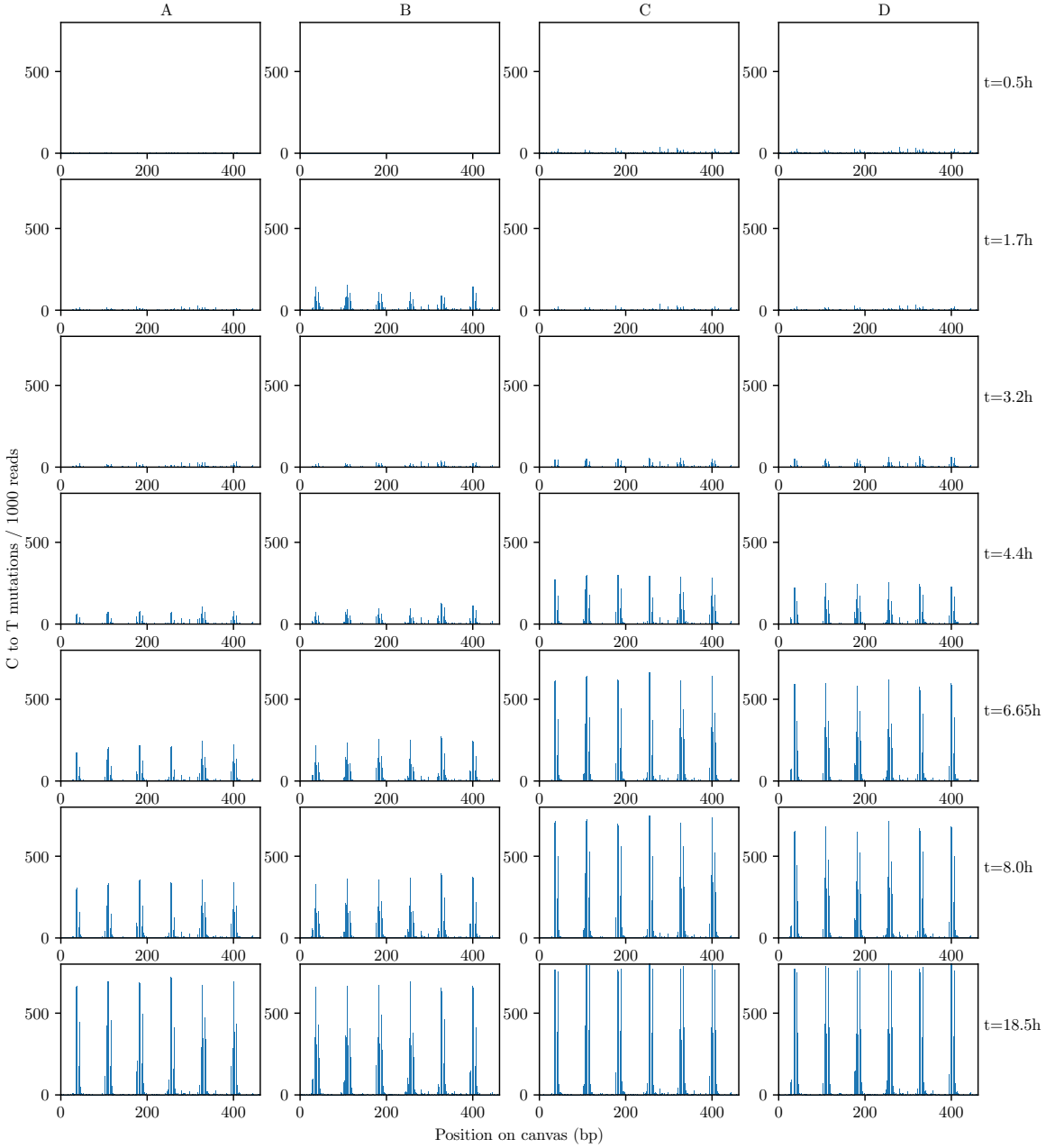



Figure S1: Representative mutation profiles of the canvas region over time. Four conditions were chosen to test the effects of base editor promoter strength, sgRNA promoter strength, and their interaction on base editing activity observed as accumulation of mutations in the canvas region. Condition A: weak base editor promoter / weak sgRNA promoter, condition B: weak base editor promoter / strong sgRNA promoter, condition C: strong base editor promoter / weak sgRNA promoter, and condition D: strong base editor promoter / strong sgRNA promoter

(a)

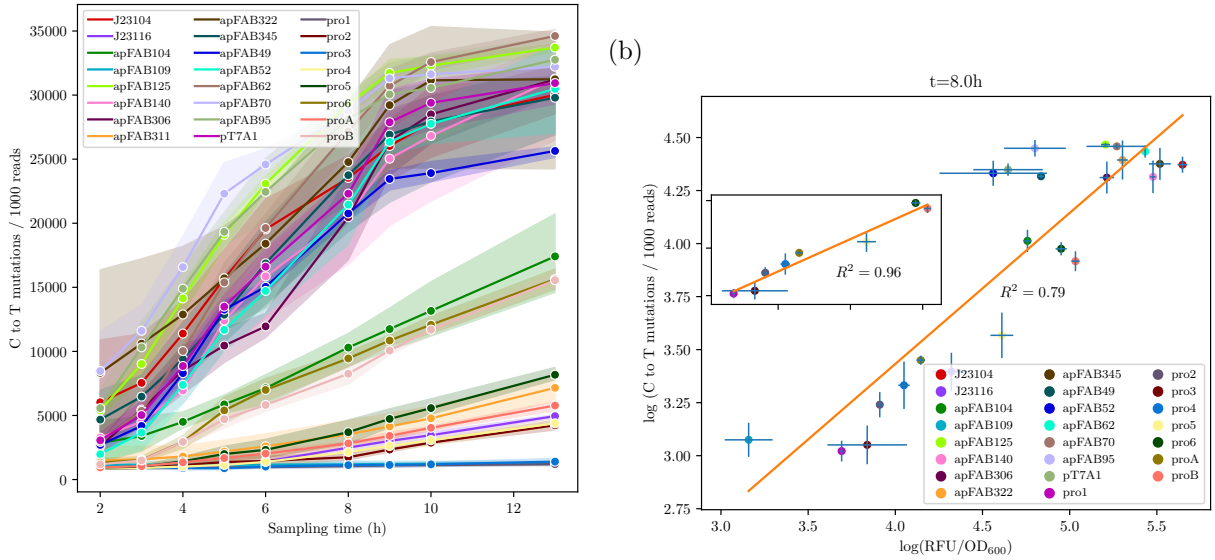



Figure S2: Validation of DHARMA performance across 24 promoters with a wide range of activities. Data were collected as in Figure 2. (a) Accumulation of C to T mutations in the canvas region of all library members within the first 12 hours of incubation. (b) The log-transformed total number of C to T mutations in the canvas region at 8h after electroporation was plotted against the corresponding log-transformed GFP fluorescence intensity for each library member. Error bar represents standard deviation. Inset: plot of insulated promoters.

(a)

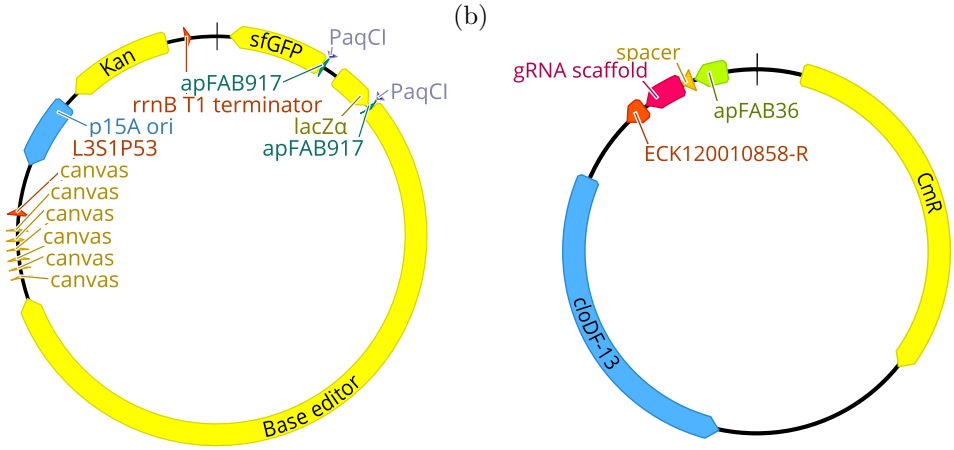



Figure S3: Maps of plasmids used in cloning and system validation. (a) Vector plasmid for library cloning. PaqCI sites flanking the lacZ$\alpha$ cassette allow scarless Golden Gate cloning of library fragments directly upstream of base editor and GFP. Blue-white screening can be used to estimate cloning efficiency and minimize background. (b) sgRNA expression plasmid. apFAB36, a strong constitutive promoter drives the expression of sgRNA targeting the repeats in canvas.

# References


[1] J. Beal, B. Teague, J. T. Sexton, S. Castillo-Hair, N. A. DeLateur, M. Samineni, J. J. Tabor, and R. a. Weiss, "Meeting measurement precision requirements for effective engineering of genetic regulatory networks," *ACS Synthetic Biology*. PMID: 35156365.

[2] N. M. Gaudelli, A. C. Komor, H. A. Rees, M. S. Packer, A. H. Badran, D. I. Bryson, and D. R. Liu, "Programmable base editing of at to gc in genomic dna without dna cleavage," *Nature*, vol. 551, no. 7681, pp. 464–471, 2017.

[3] W. Tang and D. R. Liu, "Rewritable multi-event analog recording in bacterial and mammalian cells," *Science*, vol. 360, no. 6385, p. eaap8992, 2018.

[4] B. W. Thuronyi, L. W. Koblan, J. M. Levy, W.-H. Yeh, C. Zheng, G. A. Newby, C. Wilson, M. Bhaumik, O. Shubina-Oleinik, J. R. Holt, *et al.*, "Continuous evolution of base editors with expanded target compatibility and improved activity," *Nature biotechnology*, vol. 37, no. 9, pp. 1070–1079, 2019.

[5] J. L. Doman, A. Raguram, G. A. Newby, and D. R. Liu, "Evaluation and minimization of cas9-independent off-target dna editing by cytosine base editors," *Nature biotechnology*, vol. 38, no. 5, pp. 620–628, 2020.

[6] S. Kosuri, D. B. Goodman, G. Cambray, V. K. Mutalik, Y. Gao, A. P. Arkin, D. Endy, and G. M. Church, "Composability of regulatory sequences controlling transcription and translation in escherichia coli," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, pp. 14024–14029, Aug. 2013.

[7] G. Chuai, H. Ma, J. Yan, M. Chen, N. Hong, D. Xue, C. Zhou, C. Zhu, K. Chen, B. Duan, *et al.*, "Deepcrispr: optimized crispr guide rna design by deep learning," *Genome biology*, vol. 19, no. 1, pp. 1–18, 2018.

[8] R. R. Wick, L. M. Judd, and K. E. Holt, "Deepbinner: Demultiplexing barcoded oxford nanopore reads with deep convolutional neural networks," *PLoS Comput. Biol.*, vol. 14, p. e1006583, Nov. 2018.

[9] T. F. Smith, M. S. Waterman, *et al.*, "Identification of common molecular subsequences," *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.

[10] J. H. Davis, A. J. Rubin, and R. T. Sauer, "Design, construction and characterization of a set of insulated bacterial promoters," *Nucleic acids research*, vol. 39, no. 3, pp. 1131–1141, 2011.

[11] J. Bonnet, P. Subsoontorn, and D. Endy, "Rewritable digital data storage in live cells via engineered control of recombination directionality," *Proceedings of the National Academy of Sciences*, vol. 109, no. 23, pp. 8884–8889, 2012.

[12] F. Farzadfard and T. K. Lu, "Genomically encoded analog memory with precise in vivo dna writing in living cell populations," *Science*, vol. 346, no. 6211, p. 1256272, 2014.

[13] J. Fernandez-Rodriguez, L. Yang, T. E. Gorochowski, D. B. Gordon, and C. A. Voigt, "Memory and combinatorial logic based on dna inversions: dynamics and evolutionary stability," *ACS synthetic biology*, vol. 4, no. 12, pp. 1361–1372, 2015.

[14] J. R. Rubens, G. Selvaggio, and T. K. Lu, "Synthetic mixed-signal computation in living cells," *Nature communications*, vol. 7, no. 1, pp. 1–10, 2016.

[15] S. D. Perli, C. H. Cui, and T. K. Lu, "Continuous genetic recording with self-targeting crispr-cas in human cells," *Science*, vol. 353, no. 6304, p. aag0511, 2016.

[16] F. Schmidt, M. Y. Cherepkova, and R. J. Platt, "Transcriptional recording by crispr spacer acquisition from rna," *Nature*, vol. 562, no. 7727, pp. 380–385, 2018.

[17] A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, and J. Shendure, "Whole-organism lineage tracing by combinatorial and cumulative genome editing," *Science*, vol. 353, no. 6298, p. aaf7907, 2016.

[18] K. L. Frieda, J. M. Linton, S. Hormoz, J. Choi, K.-H. K. Chow, Z. S. Singer, M. W. Budde, M. B. Elowitz, and L. Cai, "Synthetic recording and in situ readout of lineage information in single cells," *Nature*, vol. 541, no. 7635, pp. 107–111, 2017.

[19] R. Kalhor, P. Mali, and G. M. Church, "Rapidly evolving homing crispr barcodes," *Nature methods*, vol. 14, no. 2, pp. 195–200, 2017.

[20] B. C. Dickinson, M. S. Packer, A. H. Badran, and D. R. Liu, "A system for the continuous directed evolution of proteases rapidly reveals drug-resistance mutations," *Nature communications*, vol. 5, no. 1, pp. 1–8, 2014.

[21] S. Kosuri, D. B. Goodman, G. Cambray, V. K. Mutalik, Y. Gao, A. P. Arkin, D. Endy, and G. M. Church, "Composability of regulatory sequences controlling transcription and translation in escherichia coli," *Proceedings of the National Academy of Sciences*, vol. 110, no. 34, pp. 14024–14029, 2013.

[22] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, 2017.