UNIVERSITY OF CALIFORNIA, IRVINE

Engineering Mammalian Cells to Record Their Developmental Histories in DNA

DISSERTATION

submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Engineering

by

Courtney Carlson

Dissertation Committee:
Professor Chang Liu, Chair
Professor Wendy Liu
Associate Professor Tim Downing
Professor Arthur Lander

DEDICATION

To my teachers, coaches, and family members

in rural South Dakota.

Thank you for your “Midwestern nice”-ness

and for encouraging me to go after whatever I wanted.

I recall your affirming words often.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am thankful to my PhD advisor and committee chair, Chang Liu, for providing me the opportunity to work on extremely interesting and powerful research. He is a fountain of ideas for beautiful synthetic biology projects, and my time in the Liu lab has taught me to appreciate the vast potential of engineering biology and aspire to emulate his creativity. I also thank Chang for generously supplying technical advice and for always being honest when giving feedback, even in the face of disagreements. I will probably hear his voice in my head while designing experiments for the rest of my career, and I'm certain the quality of my work will be better for it.

I would also like to thank Dr. Theresa Loveless, who directly trained me in the lab, then became my co-pilot for much of my research journey. Her grittiness and talent were, and continue to be, an inspiration to me. I can't begin to recount all the technical knowledge she shared with me, and I never had more fun discussing science than in our one-on-one brainstorming sessions, where we often uncovered helpful and surprising epiphanies together.

Additionally, all members of the Liu lab deserve a thank-you, as they consistently provided technical assistance, interesting scientific discussion, and moral support. I am grateful to have spent so many hours working alongside all of them and absorbing diverse pieces of wisdom from each person.

I thank my committee members, Wendy Liu, Arthur Lander, and Timothy Downing, for bearing with my long-winded research updates and sharing thoughtful feedback on my work. I am proud to submit my dissertation with your critiques incorporated.

Also, a thank-you to Nature Chemical Biology for permission to include copyrighted text and figures from Lineage tracing and analog recording in mammalian cells by single-site DNA writing (2021).

Finally, I wish to express the copious gratitude I feel for my ride-or-die, my fiancé, Gordon Rix. He has supplied a preternatural amount of silliness in my life, which has been a

soothing tonic during the difficult times in my PhD program. I am immensely thankful for his emotional support and encouragement, and I look up to him as he is a model scientist, himself.

VITA

Courtney Carlson

EDUCATION

Doctor of Philosophy in Biomedical Engineering **2024**

University of California, Irvine

Master of Science in Biomedical Engineering **2020**

University of California, Irvine

Bachelor of Science in Chemical Engineering **2017**

South Dakota School of Mines and Technology

RESEARCH EXPERIENCE

Graduate Research Assistant **2017-2024**

University of California, Irvine

Undergraduate Research Assistant **2014-2017**

South Dakota School of Mines and Technology

NASA Jet Propulsion Laboratory

PUBLICATIONS

Carlson, C. K., Loveless, T. B., Miljsavljevic, M., & Liu, C. C. (2023) A massively parallel, *in vivo* assay of TdT mutants yields variants with altered nucleotide insertion biases. *In preparation*.

Loveless, T. B.[#], **Carlson, C. K.**[#], Dentzel Helmy, C. A., Hu, V. J., Demelo, M. C., Liang, G., … & Liu, C. C. (2023). Open-ended molecular recording of sequential cellular events into DNA. *In review at Nature chemical biology*.

Loveless, T. B., Grotts, J. H., Schechter, M. W., Forouzmand, E., **Carlson, C. K.**, Agahi, B. S., ... & Liu, C. C. (2021). Lineage tracing and analog recording in mammalian cells by single-site DNA writing. *Nature chemical biology*, *17*(6), 739-747.

Carlson, C., Singh, N. K., Bibra, M., Sani, R. K., & Venkateswaran, K. (2018). Pervasiveness of UVC 254-resistant *Geobacillus* strains in extreme environments. *Applied microbiology and biotechnology*, *102*, 1869-1887.

Singh, N. K., **Carlson, C.**, Sani, R. K., & Venkateswaran, K. (2017). Draft genome sequences of thermophiles isolated from yates shaft, a deep-subsurface environment. *Genome announcements*, *5*(22), 10-1128.

ABSTRACT

Engineering Mammalian Cells to Record Their Developmental Histories in DNA

by

Courtney Carlson

Doctor of Philosophy in Biomedical Engineering

University of California, Irvine, 2024

Professor Chang Liu, Chair

Genetically encoded DNA recorders noninvasively convert transient biological events into durable mutations in a cell's genome, thereby allowing the hidden forces that guide tissue development to be chronicled. After the developmental process of interest is complete, the cells' experiences can be read out *via* high-throughput DNA sequencing. Two types of transient events are captured with DNA recorders—the existence of progenitor cells that give rise to daughter cells (which is recorded as lineage relationships among cells by generating lineage-specific DNA barcodes), and stimuli that trigger a molecular response in the cells (which is recorded in an analog fashion by linking the accumulation of mutations to the presence of inducers of interest). This dissertation focuses on the development of two distinct, but complementary DNA recorders. First, Cell History Recording by Ordered iNsertion (CHYRON), which uses a homing guide RNA (hgRNA), CRISPR-Cas9, and a template-independent polymerase (TdT) to generate ordered insertion mutations, was developed. A large portion of this technology was already established prior to my joining the project, so this dissertation focuses only on the developments that overlapped with my efforts. Specifically, alternative nuclease components (instead of Cas9) were explored for CHYRON recording, and dose- and duration-dependent hypoxia recording was accomplished. Also, the nucleotide insertion preferences of TdT were engineered in a novel, massively parallel screen, to achieve greater lineage barcode diversity and enable recording of two different stimuli with differently-biased

TdTs. The other DNA recorder covered in this work is prime editing CHYRON (peCHYRON), which uses prime editor to generate iterative insertion mutations. Prime editor requires its edits to be precisely pre-defined in prime editing guide RNAs (pegRNAs), so pegRNA sequences that accomplish continuous insertion accumulation were carefully engineered *via* a combination of rational design and high-throughput screening. Once continuous editing was embodied by peCHYRON, we used the system for *in vitro* cell lineage tracing and molecular stimulus recording. Efforts to expand peCHYRON's ability to record a massive collection of stimuli (*i.e.*, any stimulus that can induce transcription with RNAP II) are discussed.

CHAPTER 1: Introduction to *in vivo* DNA recording

1.1 Motivation

It's an old story we've all heard: a sperm cell unites with an egg cell, a totipotent zygote is formed, and it grows into a fully formed animal comprised of many cells and specialized cell types. This is a great story—it's literally the miracle of life—but there are gaping holes in the plot. *How* does a single-celled zygote give rise to the massive coalition of cells (*e.g.*, ~37 trillion cells in the human body[1]), with distinct phenotypes (*e.g.*, ~200 cell types in humans[2]) arranged in the appropriate configuration (*i.e.*, cells, tissues, and organ systems intricately arrayed to enable collaboration among themselves), needed to recapitulate the complex body plans of animals? This question has been historically very difficult to answer due to a lack of technologies that can probe developmental processes without disrupting the underlying biology.

1.2 Historical and modern approaches to studying development

Direct observation of organismal development is the most straightforward way to track biological processes without disrupting them, but it has only been possible in rare cases. In a seminal paper by Sulston and colleagues, *C. elegans* was completely lineage-traced from the single-cell stage to a newly hatched larva *via* meticulous observation of cell-splitting patterns with an electron microscope[3]. This was possible because *C. elegans* bodies are translucent, they are comprised of a manageable number of cells (671 cells), individuals develop over a relatively short time span of 800 minutes, and cell lineage relationships are extremely invariable (so an individual doesn't need to be observed for 800 continuous minutes; rather, lineage trees from distinct individuals can be combined to create the consensus tree). Despite the strict, deterministic correspondence between cell lineage relationships and cell fate, *C. elegans* displayed unintuitive relationships among cells of differing phenotypes. For example, some

neurons originated from the mesoderm, and a few divided directly from muscles. In contrast to *C. elegans*, higher-order organisms such as mammals undergo multiple stochastic transitions, which complicates their study and suggests an important role of environmental factors in guiding their development (as opposed to lineage relationships dictating every stage)[4,5]. There have been several reported approaches to directly observe mammalian development *in utero*, (*e.g.*, by installing a window peering into the uterus of a pregnant mouse and marking the cells of interest with fluorescent proteins[6,7]), but these are constrained to short timespans. When the fetuses cease to be translucent, or when the uterine windows interfere with development, the experiments must terminate. It should also be noted that, if new techniques are developed to trace mammalian lineages for longer durations *via* direct observation, insights to the cells' experiences ("environmental factors" mentioned above, such as exposure to small molecules, cytokines, contacts with other cells, etc.) will be difficult to surmise. Fluorescent biosensors (*e.g.,* Ca^{2+} indicators[8]) may be used to indicate cellular experiences, but this approach is quite low-throughput, as the variety of events that could be tracked in parallel would be limited by the number of fluorophores that can be distinguished from each other.

Aside from direct observation, other state-of-the-art techniques for surveying development suffer from a critical drawback—they only provide a snapshot of the developing organism. They are not capable of taking continuous measurements over time because they are fundamentally disruptive to the biological processes of interest. Examples of such techniques include single-cell RNA sequencing (scRNA-seq)[9], spatial transcriptomics[10], other single-cell -omics techniques such as proteomics and epigenomics[11], imaging of fixed samples (*e.g.*, with FISH)[12], ELISA[13], and MALDI-TOF[14]. These methods all require lysing the cells, which eliminates the possibility of observing how the characterized cells will develop over time. Recently, ingenious mathematical approaches to glean temporal information from snapshots of data have come into prominence. scRNA-seq datasets, in particular, have formed the foundation for many such techniques. For instance, trajectory inference (TI) is a popular

analysis approach, with many existing algorithms for performing its computations. In essence, TI is the process of collapsing the transcriptomic data from individual cells into a reduced dimensionality space and arranging each cell along a topology that describes the development of cells over "pseudotime." A systematic comparison among 45 of the existing TI algorithms concluded they are generally accurate, but repeating the analyses with multiple different algorithms and comparing the results to cross-validate the overall topology of cellular development is strongly recommended[15]. Consequently, these analyses can be cumbersome to implement. It should also be noted that the timescale captured by trajectory inference is limited by the diversity of transcriptional states represented in the cells (*i.e.*, pseudotime can only span as wide as the assortment of cells' phenotypes allows). Another prominent technique for analyzing scRNA-seq data is RNA velocity, which infers future cell states from the observed abundance of mature mRNA and unspliced mRNAs (pre-mRNAs)[16]. A high quantity of pre-mRNAs indicates a recent increase in transcription of a gene, which is assumed to foreshadow a higher amount of spliced mRNA will exist at steady state (positive velocity). The reverse assumption is also incorporated into this model, where a low amount of pre-mRNA relative to spliced mRNA is attributed to a recent down-regulation of transcription, and the abundance of the relevant mRNA is expected to decrease in the near future (negative velocity). With this model, the transcriptomes of individual cells are subjected to dimensionality reduction (usually using UMAP or t-SNE) where cells with similar RNA velocity profiles are adjacent to each other in a vector field[17]. Closely related cell states, and the directionality of differentiation, is then inferred from these RNA velocity vector field plots. Although this model makes intuitive sense, the accuracy of the two most popular instantiations, *velocyto*[18] and *scVelo*[19], are now being called into question[16,17,20]. One critique of RNA velocity inference with *velocyto* is that it calculates velocities as a deviation from "steady-state", assuming cells in the lower and upper quantiles of the velocity vector field represent the steady-state transcription profiles. This leads to inaccurate velocity estimates if steady-state cells aren't present in the experimental data.

Also, both *velocyto* and *scVelo* use a constant splicing rate for all genes, which is known to be a crude approximation of reality. Even individual genes can exhibit variable kinetics of splicing depending on the different cell types they are expressed in. To determine which genes are satisfactorily modeled with this assumption of constant splicing rates, the curvature of the velocity vector field is consulted. No curvature constitutes an uninterpretable result, and the shape of the curve is highly gene-specific, which limits the applicability of this approach to the genes that produce curved velocity vector fields[16]. Efforts to characterize the accuracy of RNA-velocity have shown it poorly estimates the speed of cellular differentiation, except in samples with very low noise[17]. Another study on the accuracy of popular RNA velocity methods stated that a single scRNA-seq dataset from the human embryonic forebrain[18], when analyzed with *velocyto* and *scVelo*, yielded completely different results depending on the method[20]. In fact, the apparent directionality of developmental trajectories were opposites in the two sets of results. Also, similar to TI analyses, the timescale of RNA velocity is limited. The rate of RNA splicing and degradation dictates how far into the future each cell can be extrapolated.

1.3 Existing DNA recorders and their tradeoffs

To overcome the limitations of disruptive "snapshot" analysis approaches, a collection of molecular devices called *in vivo* DNA recorders (also known as "flight recorders") has rapidly expanded in recent years. The purpose of DNA recorders is to continuously log information about cells' experiences in developing tissues over time, by writing mutations to a neutral locus in the genome of the host cells. After the developmental process of interest is complete (*e.g.*, embryogenesis), the recording loci can be analyzed by Next-Generation Sequencing (NGS), and the cells' experiences can be reconstructed. "Flight recorder" is an apt analogy for these technologies, since they continuously store information which isn't read out until the series of events under study are complete. The two types of information that have been recorded with existing DNA recorders are 1) the lineage relationships among cells (which are directly recorded

by labeling cells with heritable DNA barcodes), and 2) the environmental conditions imposed on the cells (which are indirectly recorded by linking the cells' transcriptional responses to the accumulation of records). In 2016, McKenna et al. reported the first continuous DNA recorder, named Genome Editing of Synthetic Target Arrays for Lineage Tracing (GESTALT)[21]. Their platform used CRISPR-Cas9 to target a synthetic array of 10 Cas9 target sites stably integrated into the genome. Cas9-mediated double-strand breaks (DSBs) at the target sites resulted in random indels due to the error-prone, endogenous non-homologous end joining (NHEJ) pathway for DNA repair. The gradual accumulation of mutations across the 10 target sites, and the resulting combinatorial diversity of random mutations at the target sites in each cell, were used for high resolution lineage tracing in zebrafish embryos. Ultimately, hundreds to thousands of unique barcodes were generated in each zebrafish, and a lineage tree comprised of 1,323 unique barcodes was constructed for a zebrafish larva at 30 hours post fertilization (~20,000 total cells were sequenced). The efficiency of editing with GESTALT was very high, with ~60-90% of recording arrays acquiring at least one edit. However, this study presented opportunities for further development, as inter-target deletions were common among the Cas9 target sites (likely due to 2 simultaneous Cas9 cuts at distinct target sites in the same recording array, resulting in loss of the intervening targets). Since the original report of GESTALT, updated versions have been published (scGESTALT[22] and macsGESTALT[23]), as well as numerous alternative architectures.

The alternative DNA recording architectures compatible with studying mammalian tissue development can be broadly classified into 4 groups, according to the enzymes employed for mutagenesis: 1) Cas9-based methods[4,21,23–35], 2) recombinase methods[36–40], 3) base editing methods[41–45], and 4) prime editing methods[46,47]. The numbers assigned to the groups here are arbitrary. The first group, architectures that implement Cas9, benefit from high editing efficiencies, tunable editing efficiencies (mismatches in the sgRNA or altered sgRNA spacer lengths can influence the rate of editing), relatively facile implementation (few transgenes are

needed in the engineered cell lines), and they have been demonstrated for not only lineage tracing, but also environmental signal recording. However, they have several inherent limitations. For one, these systems rely on NHEJ to produce random mutations, and this often results in the generation of deletions. Although deletions can be used to uniquely barcode cell lineages (any mutation can contribute to barcode diversity, even deletions), they often result in corruption of previously-acquired mutations. Considering the example of GESTALT, it's safe to assume indels were created at individual Cas9 target sites, then were subsequently wiped out from the array of target sites when a large inter-target deletion occurred. The barcode diversity contributed by the indels in this example would be lost. Another limitation of Cas9-based systems is that they can only record information for a limited duration. Typically, a finite number of Cas9 target sites are present, so once they have all acquired an edit and are therefore no longer recognized by Cas9, editing ceases. A couple of systems endeavored to overcome this limitation by using a homing guide RNA (hgRNA)[24,34], which is a sgRNA engineered to target the gene that codes for itself. This allows Cas9 to recognize its recording locus (the hgRNA gene) even after it acquires mutations, because the target-defining sequence of the hgRNA (the spacer) always matches the sequence of the hgRNA gene. Despite the elegance of this approach, these systems gradually stop editing due to deletion of essential hgRNA elements. Another disadvantage of the Cas9-based systems is that they typically don't record information in an ordered manner, so temporal information is impossible to extract from the records. The only Cas9-based method that reliably maintains a correspondence between the order of mutations and the timing of events is CHYRON[35], which is one of the DNA recorders developed by our group and discussed in later chapters. In addition, none of the existing Cas9-based systems have demonstrated multiplexed environmental signal recording—only one signal at a time has been catalogued. The second group of DNA recorders, the recombinase-based systems, generate mutations by scrambling an arrangement of sequences between recombination sites. Because these architectures don't require DSBs for mutagenesis, they

effectively avoid deletion mutations. However, they share the other imperfections described for the Cas9-based methods: their duration of recording is limited, they don't record information in order, and they have not accomplished multiplexed signal recording. The third group of recorders, which use base editors for mutagenesis, also have the advantage of circumventing the need for DSBs. Base editor creates predictable mutations at each target site it edits (A's in the editing window are converted to G's for adenine base editor, C's are converted to T's for cytosine base editor, and other 1 to 1 conversions between different bases are made for other base editors). To establish random mutagenesis with these relatively predictable editors, many genomic loci are targeted simultaneously, so the editing outcomes (*i.e.*, which loci receive edits) is unpredictable and varies from cell to cell. These systems provide one advantage over the Cas9- and recombinase-based systems, which is that they can be used for multiplexed recording of environmental signals[44,48]. Yet, they still suffer from limited duration of recording and lack of temporal information stored in the records. Finally, the fourth clade of DNA recorders, those that leverage prime editor, is the newest group, and it overcomes many of the obstacles that have held back the other three groups of systems. We have recently established our own flight recorder that employs prime editor, called peCHYRON[46], which will be extensively discussed in later chapters. Contemporaneously with the development of peCHYRON, Choi and colleagues produced a similar system called DNA Typewriter[47]. Both peCHYRON and DNA Typewriter record mutations in order, so temporal information is stored in the spatial positions of mutations. Importantly, deletion mutations are virtually eliminated with both systems, so records are rarely corrupted. Also, both recorders have clear potential to enable multiplexed environmental signal recording. This has been indirectly demonstrated by the authors of DNA Typewriter, as they linked prime edits to distinct signals of interest[49], and envisioning how they would combine this approach with DNA Typewriter is straightforward. At the same time, we also worked toward our own method for multiplexed signal recording with peCHYRON. Our approach will be described in a later chapter. The key advantage of peCHYRON over DNA Typewriter is

that it can proceed with accruing mutations over an indefinite timespan. peCHYRON continuously generates new target sites for itself to edit, *ad infinitum*. DNA Typewriter, on the other hand, creates mutations at a pre-defined array of synthetic target sites; however, this array can be made long enough to allow many rounds of editing, so that duration of recording is not a major concern. Although the prime editing-based architectures represent several important advances in the field of DNA recording, they also have flaws that should be addressed with future developments. Namely, prime editor is notorious for exhibiting low editing efficiencies. In its current form, peCHYRON has lower overall rates of editing than DNA Typewriter, and both systems have lower efficiency than the other DNA recorders described here. Also, the special property of prime editor that is responsible for its multiplexability, compatibility with ordered recording, and ability to generate new target sites for itself, is also a source of weakness; that is, the edits installed by prime editor are precisely pre-programmed in prime editing guide RNAs (pegRNAs). This affords us unprecedented flexibility and control over the mechanism with which the recording locus propagates. However, it also requires that many transgenes be introduced into the host cells to achieve unpredictable editing outcomes (the pegRNAs don't contain random sequences, so competition among pegRNAs is used to accomplish random edits). Gene delivery of multiple pegRNAs is a crucial bottleneck when determining which applications are compatible with peCHYRON.

1.4 The pursuit of an "ideal" DNA recorder

The quest to establish a flawless DNA recorder is ongoing. As discussed above, the existing architectures are all characterized by inherent trade-offs due to fundamental mechanistic limitations. As the field advances, how can we assess when our quest for perfection is complete? Various criteria have been proposed for an ideal DNA recorder[4,46,50], which are compiled and expanded here. The perfect recorder would achieve 1) multiplexed recording of numerous signals of interest, 2) ordered recording, such that the arrangement of mutations in

the recording locus reflects the order of events the cell experienced, 3) clear delineation between adjacent records in the recording locus, 4) compatibility with simultaneous single-cell transcriptome profiling, 5) high enough recording efficiency to achieve single-cell resolution of lineages and/or environmental signals, 6) tunable rates of recording, so information is logged at a commensurate pace to the phenomena of interest, 7) vast information content, sufficient to uniquely label every cell in the host organism's body for lineage tracing, 8) minimal disruption to the biological state of the host cells, 9) durability that enables continuously operation for months or years, 10) compatibility with diverse cell types / species, 11) compatibility with simultaneous transcriptome profiling, and 12) minimal labor demands for implementation. All recorders established to date have contributed to an important and exciting leg of the journey toward creating the perfect tool for probing tissue development in live tissue. Many of the ideal characteristics outlined here have already been accomplished by the existing modalities, and the remaining challenge is to combine the strengths of the established systems. Our group has contributed to this undertaking by developing two novel DNA recorders, CHYRON and peCHYRON, with distinct strengths and weaknesses. This dissertation discusses the creation of these tools, the advantages and shortcomings of each, and provides hopeful directions to overcome the limitations of these technologies in the future.

1.5 References

1. Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Ann Hum Biol* **40**, 463–471 (2013).
2. Roy, A. L. & Conroy, R. S. Toward mapping the human body at a cellular resolution. *Mol Biol Cell* **29**, 1779–1785 (2018).
3. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev Biol* **100**, 64–119 (1983).

4. Chan, M. M. *et al.* Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).
5. Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* **453**, 544–547 (2008).
6. Kawasoe, R. *et al.* Two-photon microscopic observation of cell-production dynamics in the developing mammalian neocortex in utero. *Dev Growth Differ* **62**, 118–128 (2020).
7. Huang, Q. *et al.* Intravital imaging of mouse embryos. *Science (1979)* **368**, 181–186 (2020).
8. Mank, M. & Griesbeck, O. Genetically encoded calcium indicators. *Chem Rev* **108**, 1550–1564 (2008).
9. Saliba, A.-E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* **42**, 8845–8860 (2014).
10. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat Methods* **19**, 534–546 (2022).
11. Wang, D. & Bodovitz, S. Single cell analysis: the new frontier in 'omics'. *Trends Biotechnol* **28**, 281–290 (2010).
12. Levsky, J. M. & Singer, R. H. Fluorescence in situ hybridization: past, present and future. *J Cell Sci* **116**, 2833–2838 (2003).
13. Rennard, S. I., Berg, R., Martin, G. R., Foidart, J.-M. & Robey, P. G. Enzyme-linked immunoassay (ELISA) for connective tissue components. *Anal Biochem* **104**, 205–214 (1980).
14. Ly, A. *et al.* High-mass-resolution MALDI mass spectrometry imaging of metabolites from formalin-fixed paraffin-embedded tissue. *Nat Protoc* **11**, 1428–1443 (2016).
15. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37**, 547–554 (2019).

16. Bergen, V., Soldatov, R. A., Kharchenko, P. V & Theis, F. J. RNA velocity—current challenges and future perspectives. *Mol Syst Biol* **17**, e10282 (2021).
17. Zheng, S. C., Stein-O'Brien, G., Boukas, L., Goff, L. A. & Hansen, K. D. Pumping the brakes on RNA velocity by understanding and interpreting RNA velocity estimates. *Genome Biol* **24**, 246 (2023).
18. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
19. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* **38**, 1408–1414 (2020).
20. Gorin, G., Fang, M., Chari, T. & Pachter, L. RNA velocity unraveled. *PLoS Comput Biol* **18**, e1010492 (2022).
21. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science (1979)* **353**, aaf7907 (2016).
22. Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat Biotechnol* **36**, 442–450 (2018).
23. Simeonov, K. P. *et al.* Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. *Cancer Cell* **39**, 1150–1162 (2021).
24. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science (1979)* **353**, aag0511 (2016).
25. Schmidt, S. T., Zimmerman, S. M., Wang, J., Kim, S. K. & Quake, S. R. Quantitative analysis of synthetic cell lineage tracing using nuclease barcoding. *ACS Synth Biol* **6**, 936–942 (2017).
26. Frieda, K. L. *et al.* Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).
27. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & Van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).

28. Li, L. *et al.* A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells. *Cell* **186**, 5183–5199 (2023).
29. Bowling, S. *et al.* An engineered CRISPR-Cas9 mouse line for simultaneous readout of lineage histories and gene expression profiles in single cells. *Cell* **181**, 1410–1422 (2020).
30. Quinn, J. J. *et al.* Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science (1979)* **371**, eabc1944 (2021).
31. Spanjaard, B. *et al.* Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat Biotechnol* **36**, 469–473 (2018).
32. Jones, M. G. *et al.* Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. *Genome Biol* **21**, 1–27 (2020).
33. Kalhor, R. *et al.* Developmental barcoding of whole mouse via homing CRISPR. *Science (1979)* **361**, eaat9804 (2018).
34. Kalhor, R., Mali, P. & Church, G. M. Rapidly evolving homing CRISPR barcodes. *Nat Methods* **14**, 195–200 (2017).
35. Loveless, T. B. *et al.* Lineage tracing and analog recording in mammalian cells by single-site DNA writing. *Nat Chem Biol* **17**, 739–747 (2021).
36. Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514**, 322–327 (2014).
37. Pei, W. *et al.* Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* **548**, 456–460 (2017).
38. Pei, W. *et al.* Resolving fates and single-cell transcriptomes of hematopoietic stem cell clones by polyloxexpress barcoding. *Cell Stem Cell* **27**, 383–395 (2020).
39. Chow, K.-H. K. *et al.* Imaging cell lineage with a synthetic digital recording system. *Science (1979)* **372**, eabb3099 (2021).

40. Kotelyanskii, M., Wagner, N. J. & Paulaitis, M. E. Building large amorphous polymer structures: Atomistic simulation of glassy polystyrene. *Macromolecules* **29**, 8497–8506 (1996).

41. Farzadfard, F. *et al.* Single-nucleotide-resolution computing and memory in living cells. *Mol Cell* **75**, 769–780 (2019).

42. Hwang, B. *et al.* Lineage tracing using a Cas9-deaminase barcoding system targeting endogenous L1 elements. *Nat Commun* **10**, 1234 (2019).

43. Liu, K. *et al.* Mapping single-cell-resolution cell phylogeny reveals cell population dynamics during organ development. *Nat Methods* **18**, 1506–1514 (2021).

44. Kempton, H. R., Love, K. S., Guo, L. Y. & Qi, L. S. Scalable biological signal recording in mammalian cells using Cas12a base editors. *Nat Chem Biol* **18**, 742–750 (2022).

45. Tang, W. & Liu, D. R. Rewritable multi-event analog recording in bacterial and mammalian cells. *Science (1979)* **360**, eaap8992 (2018).

46. Loveless, T. B. *et al.* Molecular recording of sequential cellular events into DNA. *bioRxiv* (2021).

47. Choi, J. *et al.* A time-resolved, multi-symbol molecular recorder via sequential genome editing. *Nature* **608**, 98–107 (2022).

48. Tang, W. & Liu, D. R. Rewritable multi-event analog recording in bacterial and mammalian cells. *Science (1979)* **360**, eaap8992 (2018).

49. Chen, W. *et al.* Multiplex genomic recording of enhancer and signal transduction activity in mammalian cells. *bioRxiv* 2011–2021 (2021).

50. Sheth, R. U. & Wang, H. H. DNA-based memory devices for recording cellular events. *Nat Rev Genet* **19**, 718–732 (2018).

CHAPTER 2: Cell HistorY Recording by Ordered iNsertion (CHYRON) is a DNA recorder that uses Cas9, a homing guide RNA, and TdT to achieve ordered insertion mutagenesis for cell lineage tracing and analog signal recording

2.1 Introduction to CHYRON:

The first DNA recorder developed by the Liu lab was titled Cell HistorY Recording by Ordered iNsretion (CHYRON)[1]. CHYRON records information in the form of random insertion mutations (typically ~3 bp per insertion), which accumulate iteratively at a single recording locus in the genome (the CHYRON locus). The mechanism of insertion accumulation builds upon the central machinery of Kalhor et al.'s and Perli et al.'s DNA recorders—Cas9 cutting with a homing guide RNA (hgRNA)[2,3]. Therefore, CHYRON benefits from many of the same appealing features inherent to those architectures. It has high editing efficiency because it uses Cas9 for mutagenesis, it records information at a single locus (which enables facile readout *via* NGS), the rate of recording can be calibrated to the application of interest by manipulating the spacer sequence of the guide RNA, and it requires few transgenes to constitute the entire system. The key innovation of CHYRON is the use of a template-independent polymerase, terminal deoxynucleotidyl transferase (TdT) in combination with Cas9 and its hgRNA (Figure 2.1A). Owing to TdT's ability to insert random nucleotides at Cas9-mediated double strand breaks (DSBs), the mutations accrued at the CHYRON locus are primarily insertion mutations. This represents an improvement over the other Cas9-based recorders' tendency to make deletion mutations, since TdT edits the DSBs before the deletion-prone non-homologous end joining (NHEJ) pathway takes over. In a single round of editing, Cas9 cuts the locus encoding the hgRNA, and TdT makes an insertion with an unpredictable sequence composition at the DSB. The hgRNA locus is repaired, then transcribed so a new hgRNA containing the new mutation in its spacer is able to direct Cas9 back for another round of cutting, followed by another round of

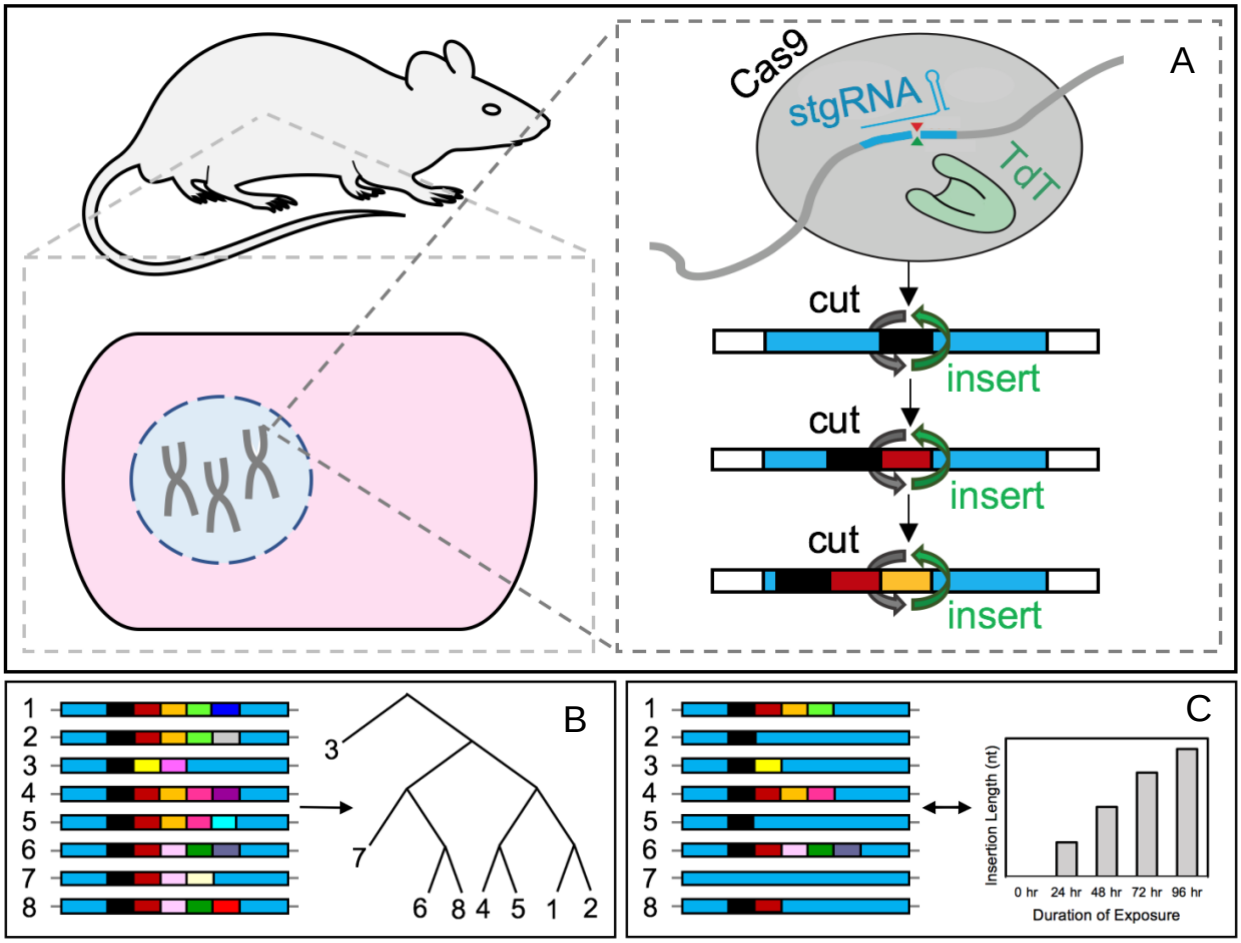


Figure 2.1. Illustration of CHYRON. **A**, At a synthetic locus in the host genome, a self-targeting guide RNA (stgRNA, a.k.a., hgRNA) directs Cas9 to cut the locus that encodes the stgRNA, itself. TdT inserts random nucleotides at the cut site. In this way, ordered insertions accumulate over time, forming a random barcode. **B**, Cell lineage tracing enabled by constitutive expression of CHYRON. The CHYRON locus is sequenced from a population of cells after proliferation. Each read is assumed to originate from a distinct cell, and lineage relationships are inferred from the barcodes. **C**, Environmental signal recording with CHYRON. In this setup, CHYRON is expressed only in the presence of a user-selected inducer; thus, the accumulation of insertions is proportional to the duration/intensity of exposure to the stimulus of interest.

TdT-mediated insertion. The process repeats until an essential component of the hgRNA gene gets corrupted due to spurious NHEJ-mediated deletions, or until the spacer of the hgRNA grows too long for efficient Cas9 cutting[2]. The insertions can serve as a long-term record of the cell lineages (if CHYRON is constitutively expressed, Figure 2.1B), or the cells' exposure to environmental cues (if CHYRON is placed under the control of a user-defined inducible promoter, Figure 2.1C). The foundation for CHYRON was established by Theresa Loveless and others prior to the initiation of my PhD[1], so this dissertation does not discuss all aspects of establishing the system. Instead, it focuses on the engineering endeavors in which I was directly

involved. This chapter will cover the experimental model for analog signal recording (various doses and durations of hypoxia exposure were recorded with CHYRON) and attempting to replace Cas9 with an alternative nuclease that possesses theoretically superior properties for long-term recording. Additionally, as a means of improving the information content of recording (improving barcode diversity for lineage tracing and enabling multiplexed signal recording), I engineered TdT to have altered nucleotide preferences. However, the discussion on TdT engineering is reserved for Chapter 3.

2.2 Methods

2.2.1 Plasmid cloning

Cloning was performed by Gibson assembly and restriction digestion and ligation cloning. All plasmids and their sequences are available at Addgene (http://www.addgene.org/browse/article/28203329/). PCR was conducted with either Q5 Hot Start High-Fidelity DNA Polymerase or Phusion Hot Start Flex DNA Polymerase (New England Biolabs (NEB)). In all cases, primers were purchased from Integrated DNA Technologies (IDT), and PCR reagents from NEB. All plasmids were prepared for transfection using the HP GenElute Midi or Mini kit (Sigma, NA0200 and NA0150).

The DNA sequence for human codon-optimized *S. pyogenes* Cas9 DNA was PCR-amplified from hCas9, which was a gift from George Church (Harvard Medical School, Addgene plasmid 41815)[4]. Using PCR and primers with overhangs containing the insertion sequence, the T2A self-cleaving sequence was inserted downstream of Cas9. To acquire the DNA sequence for TdT, cDNA from an acute lymphoblastic leukemia cell line was PCR-amplified and cloned into a pcDNA3.1 backbone, yielding Cas9-T2A-TdT. Constructs with the pcDNA3.1 backbone and Cas9 gene were transformed into XL10-Gold Ultracompetent *Escherichia coli* (Agilent, 200315).

To create hypoxia-inducible constructs, a 4xHRE-YB-TATA promoter was used to drive expression of Cas9-T2A-TdT. The 4xHRE sequence was amplified by PCR from 4xHRE_v2_YB-TATA-Gluc-CMV_dsRed30, which was a gift from Yvonne Chen, University of California, Los Angeles[5]. In addition, the sequence for the ODD was cloned to fuse with the C terminus of Cas9. The ODD sequence was amplified from HA-HIF1alpha-wt-pBabe-puro, which was a gift from William Kaelin (Harvard Medical School, Addgene plasmid 19365)[6]. This plasmid also encodes blasticidin resistance (not used in this work), the sequence for which was cloned from the pLenti CMV Blast DEST (706-1) backbone, which was a gift from Eric Campeau and Paul Kaufman (University of Massachusetts Medical School, Addgene plasmid 17451)[7].

To clone sgRNA plasmids, the spacer region of the desired sgRNA was inserted into the pSQT1313 expression plasmid, which was a gift from Keith Joung (Massachusetts General Hospital, Addgene plasmid 53370)[8], placing the sgRNA under the control of the human U6 promoter. The desired spacer region was introduced by PCR, Gibson assembly and subsequent transformation. Alternatively, a single PCR was performed on the parent plasmid to create a linear product with homologous ends. This linear piece was transformed into SS320 (Lucigen) or TOP10 *E. coli* (Thermo Fisher Scientific) to allow for recombination to yield the desired variant sgRNA plasmid.

The hgRNA constructs contained the HEK293 site 3 sgRNA cassette with a GGG (instead of GTT) sequence at the 3′ end of the spacer region and the complementary mutations in the opposite site of the hairpin[2,3]. This sequence was amplified from the sgRNA plasmid with the corresponding spacer, and the PAM-introducing mutations were present on the PCR primer. The resulting U6 promoter–hgRNA variant was cloned into a pcDNA3.1 backbone with the CMV promoter driving a puromycin-resistance gene. In addition, 750-bp regions homologous to the HEK293 site 3 locus (for $CHYRON_{20}$) or AAVS1 (for $CHYRON_{16}$) were cloned upstream and downstream of the hgRNA and selection marker region of the plasmid. The sequences of the flanks were amplified by PCR from HEK293T genomic DNA, and an EcoRI restriction site was

added on the 5′ end of the upstream flank and on the 3′ end of the downstream flank. These restriction sites allowed for linearization of the plasmid for stable integration into the genome of HEK293T cells upon transfection.

2.2.2 Cell culture and transfection

HEK293T cells were obtained from ATCC (CRL-3216) but were not otherwise authenticated. Cells were cultured in DMEM, high glucose, GlutaMAX Supplement (Gibco, 10566024), with 10% FBS (Sigma, 12306C) at 37 °C and 5% CO_2.

Transient transfections of HEK293T cells were performed by mixing DNA with FuGENE (Promega, E2311) in serum-free DMEM, at a ratio of 1 µg DNA to 3 µl FuGENE. Unless otherwise noted, the genomic site targeted for mutation was HEK293 site 3[9].

To create the 293T-CHYRON cell lines, the plasmid to integrate the hgRNA into HEK293T cells was digested with EcoRI-HF (NEB, R3101) and then purified on a silica column (Epoch, 3010). Next, 350 ng of this plasmid was mixed with 100 ng of MSP680, a plasmid expressing Cas9EQR, a gift from Keith Joung (Addgene, 65772)[10], 50 ng of a plasmid expressing an sgRNA against a sequence at HEK293 site 3 or AAVS1 that can be cut by Cas9EQR and 1.5 µl FuGENE. HEK293T cells were transfected in a 24-well dish, transformants were selected with 1–2 µg / mL puromycin (InvivoGen, ant-pr-1), and then a single colony was isolated in two rounds of dilution and colony picking. Integration into the targeted genomic locus was verified by PCR.

Samples of HEK293T and 293T-CHYRON cells used in this study, corresponding to the latest frozen stock that was used, were commercially tested for mycoplasma contamination and shown to be negative (Applied Biological Materials).

2.2.3 Hypoxia recording assay

293T-CHYRON$_{20}$ and 293T-CHYRON$_{16}$ cells were transfected in 6-well dishes with 2 μg of the 4xHRE-YB-TATA-Cas9-ODD-T2A-TdT construct. Ten hours after transfection, fresh medium supplemented with 0, 0.25, 0.5, or 1 mM DMOG (EMD Millipore Calbiochem™ #40-009) was added. Cells were collected and DNA extracted at 24 or 48 hr after DMOG addition. At 48 hr, cells were replated and retransfected 14 hr later. 14 hr after transfection, DMOG was added at the indicated concentrations, then the cells were grown for 24 hr before collection of the 72 hr timepoint, and 48 hr before collection of the 96 hr timepoint.

2.2.4 Deep sequencing library preparation of a genomic locus

Genomic DNA was isolated with a QIAamp DNA Mini Kit (Qiagen #51304). A cell pellet was lysed in a lysis buffer consisting of 20 mM EDTA, 10 mM Tris pH 8.0, 200 mM NaCl, 0.2% TritonX-100, 200 μg/μL proteinase K. The lysis reaction was incubated at 65 °C for 10 minutes and a 1:4 mixture of 7X lysis buffer (Zymo #D4036-1) to water was added and the reaction was further incubated at 65 °C. The lysis reaction was neutralized with neutralization buffer (Zymo #D4036-2) and cell debris was spun out. EconoSpin columns (Epoch #1910) were used to capture the DNA from the supernatant and a wash step with PE buffer was included before elution of the pure genomic DNA in water.

After DNA extraction, the region targeted by Cas9 was amplified by PCR. The primers contained the Illumina adapters and a 5 – 7 nt sample-specific barcode. The PCR was performed with Q5 Hot Start High-Fidelity DNA Polymerase (NEB) and the following protocol: 98°C, 1 min; (98°C, 10 s; 60°C, 30 s; 72°C, 30 s) × 35; 72 °C, 1 min. Each reaction was done with 100 ng of nucleic acid template. For the same genomic locus, each sample was normalized by signal intensity on a 0.9% agarose gel and pooled into a single mixture, which was cleaned using a NucleoSpin PCR Clean-up Kit (Macherey-Nagel #NC0389463). No size-selection was

performed, other than the exclusion of primer-sized DNA from binding to the column. 10 ng per individual sample from the pooled clean product was sent to Quintara Biosciences or FornaxBio and run on an Illumina MiSeq.

At the sequencing vendor, the libraries were purified by binding to AMPure beads (0.9 beads:1 sample) and further amplified to incorporate the TruSeq HT i5 and i7 adaptors, using Q5 High Fidelity DNA Polymerase, for 10-13 cycles. The amplified libraries were agarose gel-purified, including at least 100 bp of room around the desired bands, to avoid biasing against deletions or insertions, and then sequenced on an Illumina MiSeq using the 500-cycle v2 reagent kit (Illumina Cat # MS-102-2003).

2.2.5 Deep sequencing analysis

The sequences retrieved by next generation sequencing were first grouped to individual samples based on their barcodes. Then for each sample, associated forward and reverse reads were merged (PEAR 0.9.10[11]) and mapped to the reference sequence by the alignment algorithm implementation (Mapp) used in Perli *et al*., 2016[3], which provides a sequence of M(Match), X(Mismatch), I(Insertion) and D(Deletion) as the mapping result.

After the alignment, bad alignments (>20 mismatches or >50% deletions) were removed as the first step. Then, mismatches and inserted or deleted sequences, their positions on the reference sequence, and their frequencies were extracted. Only insertions or deletions occurring around (-10 nt to +10 nt) the cut side were kept. To remove insertions that were the result of homologous recombination repair, if longer insertions (>12nt) could map (with less than 2nt difference) to sequences of our plasmid backbones, they were filtered out. In addition, insertions longer than 15 nt were excluded, as these were found to more frequently have nucleotide biases that suggested they were TdT-independent (data not shown).

All the analyses were done in Python. The scripts are available at github.com/liusynevolab/CHYRON-NGS.

2.2.6 Statistical analyses

In all cases, biological replicates are derived from different populations of cells that were manipulated separately throughout the experiment. For technical replicates, cells were grown and manipulated, and DNA extracted, together. All procedures downstream of DNA extraction were performed separately. For the hypoxia recording experiments, statistical analysis was performed in SPSS. When analyzing percent insertions, for each sample the values of the three technical replicates were listed. For comparison of the 0.25 mM and 0.5 mM doses, the samples from each timepoint were compared by independent samples two-tailed T test (variances were equal as determined by $p>0.05$ on Levene's test). When analyzing average insertion lengths, for each timepoint and DMOG dose, a list of all insertion lengths recorded was created (for this analysis, technical replicates were pooled). Then, for each dose, whether the lengths were significantly different between timepoints was determined. For each dose, variances were unequal as determined by $p<0.05$ on Levene's test. Therefore, we performed a Welch ANOVA, followed by a post hoc Games-Howell test.

2.3 Results

2.3.1 Cas9 was replaced with a different nuclease (Cpf1) that could theoretically improve duration and information content of recording

We attempted to achieve two of the characteristics of an ideal DNA recorder, as described in Chapter 1, by engineering the nuclease component of CHYRON. Specifically, we aimed to establish punctuation between editing events (so we could interpret the precise number of edits in each record, which is useful for analog recording) and improve the duration of recording to enable the study of processes that unfold over the course of months or years.

Engineering a self-targeting version of Cpf1 (a class 2 CRISPR nuclease, meaning only one effector protein is required to cleave the target DNA[12]), was the first approach to obtaining

these improvements to CHYRON. We reasoned a self-targeting Cpf1 would be beneficial because the mechanism of hgRNA deactivation is typically deletion of the PAM in the hgRNA gene[2]. In contrast to Cas9, Cpf1 cuts its target locus distal to the PAM[13]; therefore, spurious deletions at the cut site would be less likely to affect the PAM (Figure 2.2A-B). Also, unlike the blunt-ended DSBs created by Cas9, Cpf1 generates staggered cuts with 5 nt 5' overhangs. We envisioned these staggered cuts could serve as the basis for "punctuation" between editing events. If TdT could fill in the staggered cuts in templated mode, then proceed to make random insertions, each random insertion would be flanked by the distinctive 5-nt sequence of the staggered overhang (Figure 2.2B). Also, importantly, Cpf1 has similar editing efficiency to Cas9 (often slightly lower than Cas9, but not prohibitively so); therefore, it seemed possible to reap the benefits of Cpf1 without suffering major drawbacks[14]. To this end, we tested self-targeting crRNAs for both *Acidaminococcus sp.* Cpf1 and *Lachnospiraceae bacterium* ND2006 Cpf1. Endowing the crRNAs with self-targeting ability followed the same approach as making self-targeting Cas9 sgRNAs[2,3]. Briefly, the protospacer adjacent motif (PAM) for Cpf1, which is 5'-TTTN-3', was inserted into the crRNA gene, in the appropriate position adjacent to the spacer sequence, in region of the gene that encodes the 5' handle of the crRNA (Figure 2.2B). On either side of the position where the PAM was inserted, the wildtype crRNA scaffold has thymines, so introducing a PAM created the sequence 5'-TTTTT-3', which is a long enough stretch of homopolymeric thymine to cause ~95% efficient transcriptional termination by RNA polymerase III[15]. Thus, to enable transcription of full-length crRNAs, the thymine at the -5 position was converted to a different nucleotide. We mutated T(-5) rather than T(-1) because previous work[16] has shown that the U(-1)•U(-16) base pair in the crRNA is essential for Cpf1-mediated DNA cleavage, as Cpf1 pairs with its crRNA via base-specific interactions at this position. For each of these mutated nucleotides, compensatory mutations were introduced into the hairpin of the 5' handle, to ensure the native secondary structure was preserved. These self-

targeting crRNAs were assayed in HEK293T cells *via* transient transfection of the crRNAs plus the appropriate Cpf1, then sequencing the crRNA genes 3 days post-transfection.

Illumina sequencing revealed the self-targeting crRNAs did not produce detectable editing. The failure to edit is suspected to be caused by the presence of a polythymidine tract in the crRNA. Final designs had tetrameric thymidines (5'-TTTT-3') in the 5' handle, which likely caused enough transcriptional termination (up to 75% termination[15]) to substantially reduce editing rates. The position of this putative transcription termination signal would facilitate transcription of the crRNA's scaffold, but the spacer would be absent. Therefore, ~75% of crRNAs in the cell would be capable of binding with Cpf1, but incapable of editing the target locus. Competition between these truncated crRNAs and the ~25% fully transcribed crRNAs could be sufficient to reduce editing rates to the point of no detection. The work by Gao and colleagues that characterized the exact length of thymidine homopolymers needed to terminate transcription with RNA pol III was published at approximately the same time we began this Cpf1 engineering project; therefore, we were unaware of the extent to which this polythymidine tract could inhibit crRNA expression when we conducted this work.

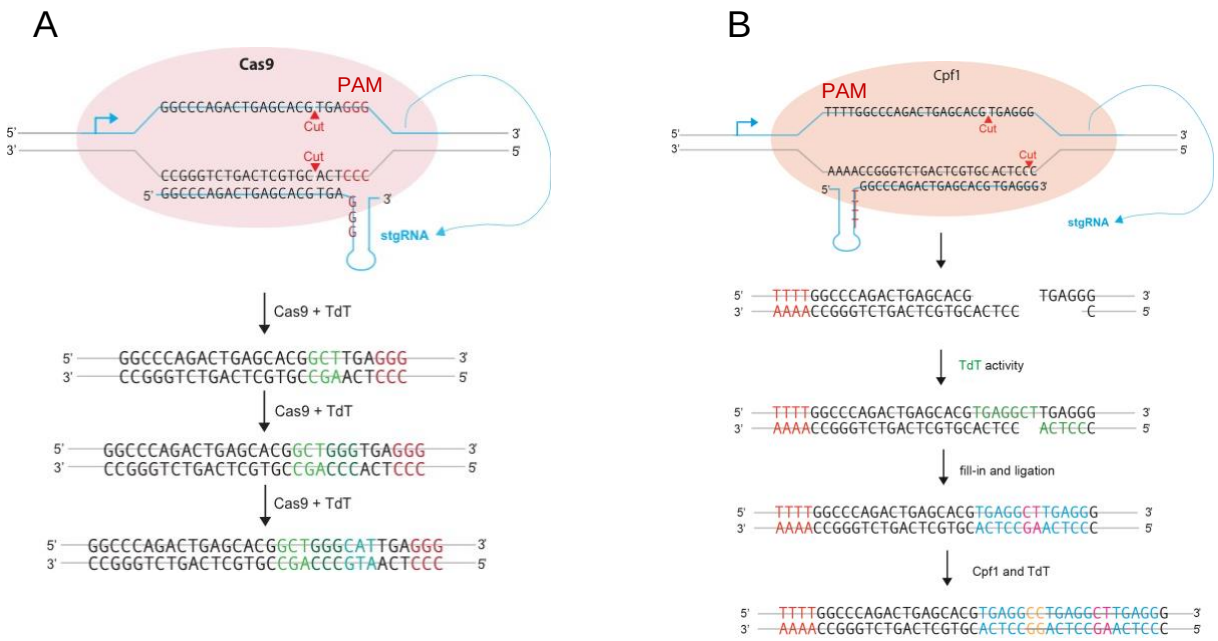


Figure 2.2. Comparison of Cas9 and Cpf1 in CHYRON editing scheme. **A,** CRISPR-Cas9, the original nuclease used for CHYRON recording, creates blunt-ended double-strand breaks (DSBs) 3 bp upstream of its PAM. The proximity of the DSBs to the PAM places the PAM in jeopardy of being corrupted if a deletion occurs. After TdT creates insertions, Cas9 cuts again at the exact same site. Because the DSBs are blunt, TdT-mediated insertions accumulate without clear delineation between editing events. **B,** Cpf1 cleaves its target site distal to the PAM, which reduces the risk of the PAM being corrupted. Also the Cpf1-mediated DSBs have a 5 nt 5' overhang. We hypothesized TdT would fill in the overhang in templated mode, then insert random nucleotides. This would create repeats of the 5-bp overhang sequence on either side of each TdT insertion. These repeats could be used as "punctuation" between each edit.

As an alternative to using Cpf1 in a self-targeting configuration, we also assessed the feasibility of using Cpf1 for repeated editing at a recording locus, while expressing the crRNA from a separate locus. Because Cpf1 cuts its target site distal to its seed region, the mutations at the cut site might not significantly hinder target recognition and cutting. In fact, only 1 nt of the target site would be expected to differ from the original target sequence after TdT-mediated insertion mutagenesis (Figure 2.2B). To see if such a 1 nt mismatch would interfere with editing, Cpf1 was assayed with a regular (non-self-targeting) crRNA that diverges from its target site in the last nt of the spacer. Cpf1 and its crRNA were co-transfected with TdT into HEK293Ts, relying on TdT's recognition of the Ku complex at a DSB to recruit TdT to the cut site. To improve the chances of TdT interacting with the Cpf1-generated DSB, we also attempted recruiting TdT to a locus adjacent to the cut site, using a fusion protein of dCas9-XTEN-TdT.

Encouragingly, the activity of Cpf1 when guided by a crRNA with a 1 nt mismatch was nearly identical to that of Cpf1 with a crRNA that perfectly matches its target (data not shown). However, insertion mutations were very rare, indicating the staggered cuts created by Cpf1 are not a suitable substrate for TdT. Ultimately, Cpf1 delivered on its promise of cutting its target site in a manner that would increase the durability of recording (*i.e.*, mutations would accrue distal to the essential sequence motifs within the target, and Cpf1 could tolerate mutation accumulation even if performing target recognition in a non-self-targeting fashion); however, its inability to productively partner with TdT is a critical issue. To unlock the beneficial features of Cpf1 for CHYRON, future work needs to focus on improving synergy between Cpf1 and TdT.

2.3.2 Hypoxia recording with CHYRON

We established CHYRON's analog signal recording capability in a proof-of-principle setup with a synthetic hypoxia-responsive promoter driving expression Cas9 and TdT (while the hgRNA was constitutively expressed). We used the 4xHRE-YB_TATA promoter[5] to express a polycistronic cassette comprised of Cas9-ODD-T2A-TdT, where ODD is an oxygen-dependent degron, and the T2A linker is a self-cleaving peptide. With the hypoxia responsive promoter controlling transcription of this cassette, the accumulation of CHYRON insertions would progress whenever hypoxic conditions were present. The 4xHRE-YB_TATA promoter was selected because it has been previously shown to have minimal basal expression and robust induction[5]. We chose to express Cas9 and TdT in a single cassette because both DNA-editing enzymes should be expressed at the same time to promote the chances of both enzymes collaboratively editing the recording locus (whereas expression of Cas9 without TdT would result in unwanted deletions). We tested the efficacy of this approach in two different CHYRON cell lines—one that constitutively expresses a hgRNA with a 20 nt spacer sequence, $CHYRON_{20}$ (the optimal length for Cas9 editing efficiency), and another that expresses a hgRNA with a 16 nt spacer, $CHYRON_{16}$ (each cell has low editing efficiency initially, but more edits can accumulate before

the hgRNA becomes too long to perpetuate the recorder). We transfected both cell lines with the 4xHRE-YB-TATA-Cas9-ODD-T2A-TdT construct, then exposed them to differing concentrations of a chemical hypoxia mimic (DMOG), for differing durations. In both cell lines, there was a clear increase in editing if the cells were exposed to DMOG (compared to cells receiving no DMOG treatment) (Figure 2.3A-C). In the case of $CHYRON_{20}$, the dosage of DMOG had a direct correlation with the overall extent of editing in the population of cells (Figure 2.3A). For $CHYRON_{16}$, however, DMOG concentration did not have a statistically significant effect on the percentage of loci bearing an edit (Figure 2.3B). We believe the lack of dose dependence is due to the low editing efficiency inherent to a hgRNA with a 16 nt spacer, which reduces the dynamic range. We assessed if a different metric, the average insertion length, in the $CHYRON_{16}$ population could discern between different DMOG doses, but again the results were not significant (Figure 2.3C). We speculate this is because CHYRON has a binary "off" or "on" behavior in individual cells. Therefore, dose of DMOG determines the probability that CHYRON will be turned "on" in a given cell (hence the change in percent edited loci with respect to dosage), but does not affect the rate of insertions accruing in individual cells. Although the $CHYRON_{16}$ cell line could not reliably report on different doses of DMOG, it was able to discern between different durations of exposure *via* its average insertion lengths (Figure 2.3C). In the end, this hypoxia recording demonstration showed that CHYRON can carry out digital recording (presence vs. absence of hypoxia is distinctly encoded) and analog recording (different doses of hypoxia were successfully recorded with $CHYRON_{20}$, while durations were captured with $CHYRON_{16}$). We acknowledge the dynamic range of the analog recordings leave room for improvement, but we expect this could be augmented with future developments to the system.

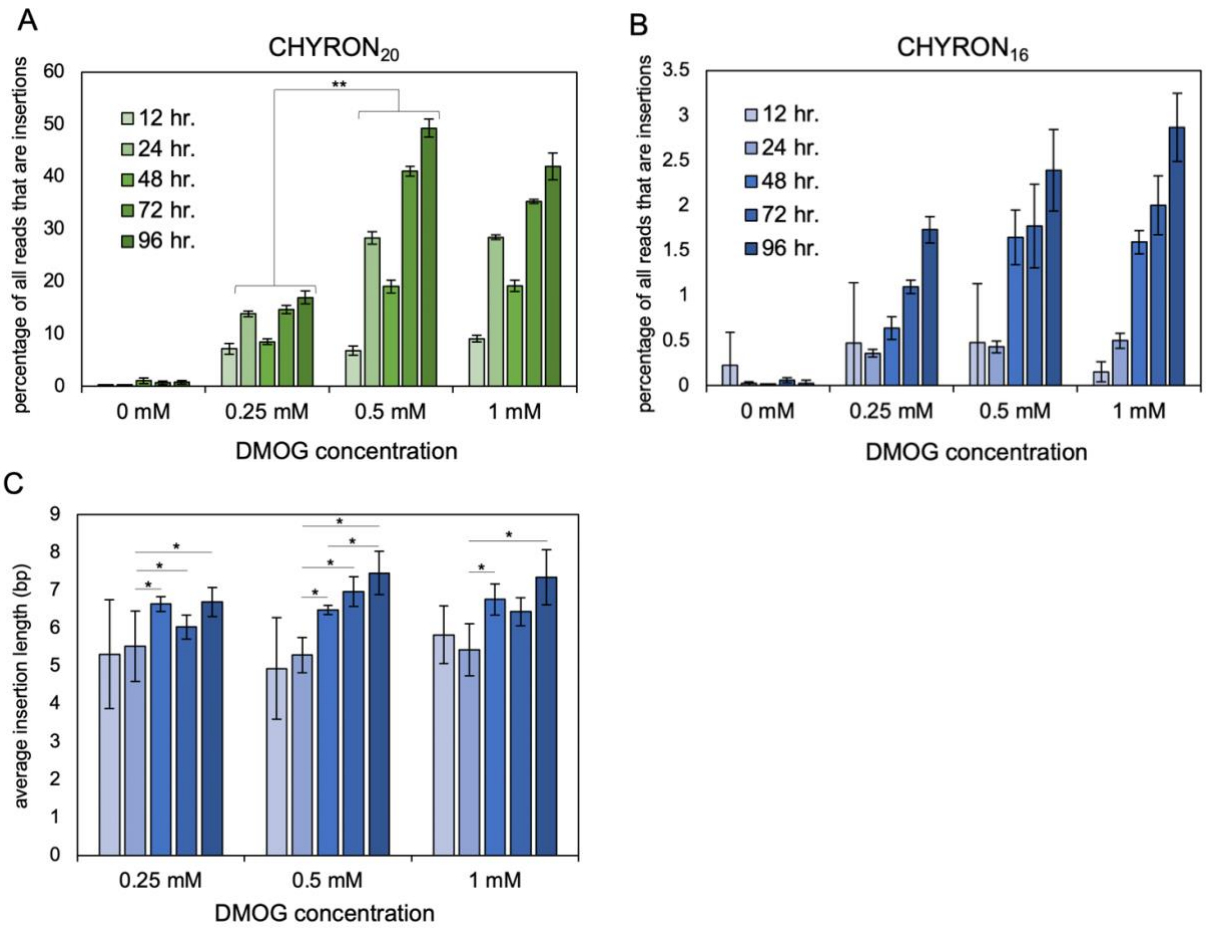


Figure 2.3. A hypoxia-inducible Cas9 and TdT construct was used to record dose and duration of exposure to a chemical hypoxia mimic called DMOG. **A,** A population of $CHYRON_{20}$ cells were transfected with the hypoxia-inducible CHYRON construct. Different doses of DMOG were added to each culture, and the samples were collected and DNA sequenced after the indicated amount of time. Insertion mutations accumulated in a dose-dependent manner. Each data point is the average of three technical replicates. Error bars=±stdev. For the samples marked **, the insertions observed for different doses were significant by two-tailed T test. For the 24 hr. timepoint, t-value=19.79, p<0.001; for the 48 hr. timepoint, t-value=13.484, p<0.001; for the 72 hr. timepoint, t-value=36.002, p<0.001; for the 96 hr. timepoint, t-value=26.404, p<0.001. **B,** A $CHYRON_{16}$ cell line accrued insertions at a lower, rate, without apparent dose dependence. $CHYRON_{16}$ was subjected to the same experimental protocol as $CHYRON_{20}$. **C,** The 293Ts expressing $CHYRON_{16}$ recorded the duration of DMOG exposure *via* changes in average insertion lengths. Data points represent the average of three technical replicates. Error bars=±stdev. All timepoints at each dose were significantly different by Welch analysis of variance (ANOVA). For the 0.25 mM dose, F=17.659, p<0.001; for the 0.5 mM dose, F=18.463, p<0.001; for the 1 mM dose, F=7.461, p<0.001. Pairs of samples marked * were significantly different according to post hoc Games-Howell test (p<0.005).

2.4 Discussion

The original CHYRON system, which used a hgRNA plus Cas9 to cut the recording locus, and wildtype TdT to write insertions at the resulting DSBs, was shown in the Loveless et al. publication to effectively shift DNA repair outcomes from deletions (which are the predominant

outcome for Cas9 alone) to insertions (which occur ~80% of the time when TdT is co-expressed with Cas9)[1]. While it wasn't expounded upon in this document, it is important for context to note this technology was used to successfully record lineage relationships among populations of cells in a proof-of-principle, cell culture-based experiment in the Loveless et al. publication.

This chapter described my most significant contributions to establishing CHYRON as one of the most versatile DNA recorders to date (capable of both high information content recording for cell lineage tracing, and analog signal recording). Our engineering approach was two-pronged: for one, we attempted to establish Cpf1 as the nuclease for CHYRON, which was expected to improve the durability of recording and clarity of records. Secondly, we replaced the promoter driving expression of the recording machinery, such that accumulation of records could be linked to the presence of an inducer of interest.

The roll-out of Cpf1 as CHYRON's new nuclease was ultimately unsuccessful, but reasonable, addressable explanations exist for the observed lack of CHYRON recording. In the self-targeting configuration of Cpf1, we selected Cpf1 genes from *Acidaminococcus sp.* and *Lachnospiraceae bacterium* ND2006 because they have comparable efficiencies to Cas9[14], and they are the two most commonly used variants of Cpf1[17], and therefore the best characterized. However, the PAM requirement for these variants (5'-TTTN-3')[18] presented challenges to designing a self-targeting crRNA. Placing three contiguous thymidines adjacent to the target site necessitated mutagenizing the scaffold region of the crRNA. The structure of the scaffold is important, since it mediates binding of the crRNA to the Cpf1 protein by slotting into the positively charged groove between Cpf1's WED and RuvC domains[19]. Therefore, for every mutation we installed to constitute the TTTN PAM, we also introduced compensatory mutations to conserve the overall structure of the scaffold. We took a conservative approach, mutating the minimal number of nucleotides, but this resulted in a tetramer of thymidines in the self-targeting crRNA gene. This 5'-TTTT-3' was found (around the same time we began our Cpf1 crRNA engineering efforts) to be a potent transcription termination signal for RNA polymerase III

(RNAPIII) if expressed by a U6 promoter[20]. Other RNAPIII promoters yield far less termination with the same 5'-TTTT-3' sequence (27% and 18% termination efficiency for p7SK and pH1, respectively)[20]. Therefore, it may be fruitful to proceed with the self-targeting crRNA sequence we already designed, but with a different promoter. An alternative promising approach would be to redesign the self-targeting crRNA, but implement new variants of *Acidaminococcus sp.* and/or *Lachnospiraceae bacterium* ND2006 that have been engineered to have reduced PAM specificities[17]. The new PAM requirements are 5'-TYCV-3' and 5'-TATV-3'. These variants would allow Cpf1 to target the gene for its own crRNA without needing to substitute so many bases in the crRNA scaffold, and without introducing any transcription termination signal.

Aside from establishing a self-targeting Cpf1, using Cpf1 to repeatedly edit the recording locus in a non-self-targeting manner is also promising. How long this design can perpetuate editing before the mismatches between the crRNA and the recording locus become intolerable has yet to be characterized, but our initial experiment showing a 1 nt mismatch has no obvious effect on efficiency is a good omen. Presently, the critical stumbling block of this approach is that TdT doesn't use the staggered-cut DNA with a 5' overhang as an efficient substrate for insertion mutagenesis. This could be ameliorated in the future by experimenting with different sequences at the cut site (*i.e.*, TdT prefers single-stranded 3' ends as its substrate, so more A/T-rich sequences in the duplex could help by promoting DNA melting and TdT latching onto the exposed 3' end), engineering a blunt-cutting Cpf1, or employing additional polymerases (*e.g.*, Klenow fragment) to blunt the staggered cuts prior to TdT-mediated insertion. Alternatively, TdT could be engineered to work more efficiently with staggered cuts. Detailed engineering approaches for this are discussed in Chapter 3.

2.5 References

1. Loveless, T. B. *et al.* Lineage tracing and analog recording in mammalian cells by single-site DNA writing. *Nat Chem Biol* **17**, 739–747 (2021).
2. Kalhor, R., Mali, P. & Church, G. M. Rapidly evolving homing CRISPR barcodes. *Nat Methods* **14**, 195–200 (2017).
3. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science (1979)* **353**, aag0511 (2016).
4. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science (1979)* **339**, 823–826 (2013).
5. Ede, C., Chen, X., Lin, M.-Y. & Chen, Y. Y. Quantitative analyses of core promoters enable precise engineering of regulated gene expression in mammalian cells. *ACS Synth Biol* **5**, 395–404 (2016).
6. Yan, Q., Bartz, S., Mao, M., Li, L. & Kaelin Jr, W. G. The Hypoxia-Inducible Factor 2 α N-Terminal and C-Terminal Transactivation Domains Cooperate To Promote Renal Tumorigenesis In Vivo. *Mol Cell Biol* **27**, 2092–2102 (2007).
7. Campeau, E. *et al.* A versatile viral system for expression and depletion of proteins in mammalian cells. *PLoS One* **4**, e6529 (2009).
8. Tsai, S. Q. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat Biotechnol* **32**, 569–576 (2014).
9. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
10. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).

11. Tanida-Miyake, E., Koike, M., Uchiyama, Y. & Tanida, I. Optimization of mNeonGreen for Homo sapiens increases its fluorescent intensity in mammalian cells. *PLoS One* **13**, e0191108 (2018).

12. Tang, Y. & Fu, Y. Class 2 CRISPR/Cas: an expanding biotechnology toolbox for and beyond genome editing. *Cell Biosci* **8**, 1–13 (2018).

13. Safari, F., Zare, K., Negahdaripour, M., Barekati-Mowahed, M. & Ghasemi, Y. CRISPR Cpf1 proteins: structure, function and implications for genome editing. *Cell Biosci* **9**, 1–21 (2019).

14. Kleinstiver, B. P. *et al.* Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nat Biotechnol* **34**, 869–874 (2016).

15. Gao, Z., Herrera-Carrillo, E. & Berkhout, B. Delineation of the exact transcription termination signal for type 3 polymerase III. *Molecular Therapy-Nucleic Acids* **10**, 36–44 (2018).

16. Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759–771 (2015).

17. Gao, L. *et al.* Engineered Cpf1 variants with altered PAM specificities. *Nat Biotechnol* **35**, 789–792 (2017).

18. Yamano, T. *et al.* Crystal structure of Cpf1 in complex with guide RNA and target DNA. *Cell* **165**, 949–962 (2016).

19. Yamano, T. *et al.* Crystal structure of Cpf1 in complex with guide RNA and target DNA. *Cell* **165**, 949–962 (2016).

20. Gao, Z., Herrera-Carrillo, E. & Berkhout, B. Delineation of the exact transcription termination signal for type 3 polymerase III. *Molecular Therapy-Nucleic Acids* **10**, 36–44 (2018).

CHAPTER 3: Engineering terminal deoxynucleotidyl transferase (TdT) to have altered nucleotide preferences *via* a high throughput, pooled screen in HEK293T cells

3.1 Introduction:

Terminal deoxynucleotidyl transferase (TdT) is one of the only known polymerases that can use single-stranded nucleic acids as its substrate and append nucleotides in a template-independent manner[1,2]. The product of this unique activity is single-stranded DNA (or RNA) with random sequence composition. In nature, TdT's erratic nucleotide incorporation plays an essential role in the adaptive immune systems of vertebrates—TdT is responsible for diversifying T-cell receptors during V(D)J recombination, resulting in a vast repertoire of distinct antigen receptors. Thus, randomness is an essential part of its job.

Because of its ability to extend single-stranded nucleic acids, molecular biologists have plucked TdT out of its natural niche and thrust it into a wide range of biotechnological applications requiring *de novo* DNA or RNA synthesis. For some applications, such as barcoding cell lineages *in vivo* with random insertion sequences[3,4] (*e.g.*, with CHYRON), or generating libraries of variable-sequence polynucleotides to bind protein targets[5], the unpredictable nature of TdT-generated sequences is desirable. In most cases, however, the randomness of the inserted nucleotides is a liability that must be reined in. A straightforward approach to controlling the products of TdT's synthesis is pre-defining the pool of nucleotides provided to TdT, so it has no choice but to construct oligomers from the limited set of building blocks. This is appropriate for applications that can tolerate variable lengths of the resulting nucleic acids, such as biosensing[6–13], end-labelling of DNA or RNA with natural or unnatural/modified nucleotides[14,15], and DNA data storage[16]. Aside from manipulating the dNTPs, other components of the reaction mixes have been fine-tuned to stifle TdT's zest for random extension (*e.g.*, TdT's nucleotide preferences can be influenced by adjusting the

composition of divalent cation cofactors, and this approach has been leveraged for DNA data storage[17], and others have co-opted a dNTP-degrading enzyme called apyrase to limit the length of TdT-synthesized ssDNA during DNA data storage[16]). While these approaches have yielded several practical applications and many exciting proofs of principle, it is worth noting that the fundamental drawbacks of TdT have sparsely been addressed at the root of the problem (*via* protein engineering). To our knowledge, the only protein engineering campaigns for TdT have pertained to the most extreme case of controlling TdT's activity (*i.e.*, oligonucleotide synthesis, which requires stepwise addition of precise nucleotides[18–21]). However, we believe many of the existing TdT applications could benefit, and novel technologies could be enabled, with TdT variants tailored to each application. For instance, TdTs with altered nucleotide preferences would be an asset to *in vivo* DNA recording, potentially increasing the information content of random barcodes for cell lineage tracing and enabling multiplexed signal recording (discussed later in detail). Also, TdT variants possessing faster kinetics could boost the efficiency of oligo synthesis with TdT (*e.g.*, for one of the state-of-the-art methodologies, dTTP requires 2x longer coupling times than the other dNTPs, presenting a ripe opportunity for optimization[18]). Given the massive demand for synthetic oligonucleotides, even modest gains in extension kinetics would have a dramatic impact on the overall output of these platforms.

Protein engineering efforts to create bespoke TdTs have likely been limited by the laborious and generally low-throughput nature of assaying TdT variants *in vitro*. In a herculean feat, Lu and colleagues individually expressed, purified, and assayed 2,180 TdT mutants to establish a variant with a catalytic cavity that can accommodate 3' O-blocked dNTPs for oligo synthesis[20]. Toward improving TdT's thermostability to facilitate oligo synthesis at elevated temperatures, Chua et al. developed a FRET-based assay for TdT activity[19]. This relatively high-throughput approach allowed 10,882 variants to be screened in *E. coli* lysates. Still, each variant required individual expression, lysis, and assaying, necessitating an impressive amount of time and resources for their work. Also, while the FRET-based assay was a powerful tool for

their screen, it requires Cy5-labeled nucleotides to be the substrates for TdT incorporation, and therefore is not compatible with efforts to engineer TdT's compatibility with other nucleotides. Others have successfully used rational design to devise, then assay a smaller number (<5) of TdT variants for oligo synthesis / increased thermostability[18,21]. In all these cases, the engineering efforts were successful. However, the prospect of repeating similar resource-intensive, large-scale assays for other desired TdT properties is daunting, and opportunities for rationally engineering TdT to obtain new properties in a manner that won't necessitate screening thousands of variants may be rare.

In the present study, we developed a novel, massively parallel assay for TdT that enables pooled screening of tens of thousands of variants simultaneously in HEK293T cells. We demonstrated the efficacy of this approach by engineering TdT variants with altered nucleotide insertion preferences (A/T biases), since different compositions of insertions are useful for *in vivo* DNA recording with our established recorder, CHYRON[22]. In the process of creating altered-bias TdTs, we also discovered variants with altered average insertion lengths, which we envision having utility for the previously mentioned applications of TdT. We generated 3 site saturation mutagenesis libraries of TdT with 27,783 possible amino acid sequences (98,304 DNA sequences). 83.4% of the possible amino acid sequences were assayed in a sufficient number of independent instances to assess their nucleotide preferences and average insertion lengths with a reasonable level of confidence from the pooled assay. We intend for the resulting database of characterized TdT variants to be an open resource for others developing biotechnological applications of TdT. To demonstrate the reliability of the pooled assay results, we selected 75 TdT variants spanning a range of nucleotide insertion preferences and average insertion lengths, then assayed them individually. This confirmed the high-throughput assay results are reproducible *in vivo*. In addition, we explored the reproducibility of library members' nucleotide preferences when assayed *in vitro*. We observed the A/T biases were conserved

regardless of the *in vitro* or *in vivo* nature of the assay. Finally, we demonstrated the utility of altered-bias TdTs for DNA recording in live mammalian cells.

3.2 Methods:

3.2.1 Clonal plasmid construction

The only sgRNA needed in this study was one targeting the HEK3 sequence (Supplementary Table 7). The HEK3 sequence was used as the Cas9 target site in the high-throughput screen (it was encoded in the lentivirus cassette to keep the TdT library member gene and Cas9 target site on the same molecule), and the endogenous HEK3 locus was targeted during individual assays of TdT library members. It was cloned *via* Gibson Assembly, adding the appropriate spacer sequence into the pSQT1313 vector (Addgene #53370), which was a gift from Keith Joung[23]. After confirming faithful assembly *via* Sanger sequencing (Genewiz) or whole-plasmid sequencing (Primordium Labs), the sgRNA plasmid was then hp-miniprepped (GenElute HP Plasmid Miniprep Kit, NA0150) to produce high-purity plasmid stocks for downstream transfections in HEK293T cells.

TdT library members individually assayed in HEK293Ts were cloned by first inserting wildtype TdT into a pcDNA3.1 expression plasmid (this will be made available on Addgene), then adding point mutations *via* Gibson or Golden Gate Assembly with the mutations encoded in junctions between fragments. Assembled plasmids were confirmed *via* Sanger sequencing (Genewiz) or whole-plasmid sequencing (Primordium Labs), then hp-miniprepped (GenElute HP Plasmid Miniprep Kit, NA0150) to produce high-purity plasmid stocks for downstream transfections.

For TdT library members that were installed into the CHYRON architecture, a piggyBAC expression vector from the Mammalian Toolkit[24] was used. The Mammalian Toolkit was a gift from Hana El-Samad (Addgene kit #1000000180). For continuous propagation of the CHYRON

locus, the pCAG promoter was used (Addgene #123700, also a Mammalian Toolkit part plasmid) to drive expression of a Cas9-T2A-TdT cassette, which was cloned from one of our previously described plasmids (Addgene #132667)[22]. Either Gibson Assembly or Golden Gate Assembly was used to first make the piggyBAC-pCAG-Cas9-T2A-TdT plasmid, then point mutations were subsequently added to TdT to constitute the variants of interest. For dual signal recording, the same piggyBAC expression vector and Cas9-T2A-TdT fragments were used. However, the promoter was either 4xHRE_YB TATA[25] (Addgene #117399) for hypoxia recording or 7xTCF/LEF[26] (Addgene #12456). M50 Super 8x TOPFlash containing the 7x TCF/LEF promoter was a gift from Randall Moon, and 4xHRE_YB TATA-sfGFP-CMV_dsRed was a gift from Yvonne Chen. Also, an oxygen-dependent degron from our previously described plasmid (Addgene #132667) was appended to Cas9 in the hypoxia-recording cassette. All pertinent plasmids from this study will be made available on Addgene.

For *in vitro* assays, a truncated version of the gene for wildtype human TdT was codon-optimized for expression in *E. coli* B (Integrated DNA Technologies, https://www.idtdna.com/CodonOpt) and synthesized as a gBlock from IDT. The N terminus was excluded, as we have found the BRCT domain[27] compromises expression, and it is unnecessary for *in vitro* characterization. In the final design, only residues 125-509 were included. The TdT gene was cloned into a pET28a expression vector with an N-terminal 10xHis tag, linker, and protease cleavage site. The wildtype TdT plasmid was used as a template for targeted mutagenesis to produce plasmids for each variant of interest, and it will be made available on Addgene.

3.2.2 Site saturation mutagenesis library cloning and lentivirus production

Human TdT was amplified from one of our previously reported plasmids (Addgene #163643) and cloned into a pCMV Blast Dest lentiviral expression vector (Addgene #17451) *via* Gibson Assembly. At the same time, pre-existing SalI and PstI restriction sites were removed *via*

introduction of synonymous point mutations, the ampicillin resistance gene in the lentiviral backbone was replaced with kanamycin resistance, puromycin resistance was added for later selection of transduced cells, and a sfGFP gene was added for virus titering. In addition, new SalI and PstI restriction sites were added *via* synonymous point mutations, to enable downstream cloning steps. From this wildtype TdT template plasmid, libraries 1 and 2 were cloned *via* restriction digestion with SalI-HF and PstI-HF followed by ligation with T4 DNA Ligase, and library 3 was cloned *via* Golden Gate Assembly with Esp3I. All enzymes were purchased from New England Biolabs. For all libraries, saturation mutagenesis was achieved through 2 consecutive PCRs on the TdT gene. In the first PCR, most of the TdT gene was amplified, from the promoter driving TdT expression to the last nucleotide upstream of the NNK'd codons. This PCR product was digested with DpnI to remove wildtype template plasmid, then purified *via* extraction from an agarose gel (Epoch Life Sciences All-In-One DNA Only Spin Columns). The cleaned PCR was used as the template for a second PCR that introduced NNK degenerate codon sequences as primer overhangs (primers were synthesized by Integrated DNA Technologies, with hand-mixed degenerate bases harboring 1:1:1:1 composition of A:C:G:T for N and 0:0:1:1 for K sites in the NNKs). These primers also introduced restriction sites for the downstream plasmid assembly. The PCR products were cleaned directly on Epoch Life Sciences All-In-One DNA Only Spin Columns (no gel extraction). For libraries 1 and 2, the TdT gene inserts and plasmid backbones were digested in separate reactions with 1 ug of DNA per 50 uL digestion reaction, 100 U of SalI-HF, 20 U of PstI-HF, and Cutsmart buffer (New England Biolabs). For the plasmid backbones, a miniprep of the wildtype TdT plasmid was digested directly, and CIP (New England Biolabs) was added to the digestion to prevent empty vector self-ligation. The digestion was incubated at 37°C for 1 hour. Digestion products were extracted from agarose gels, then the inserts and backbones were ligated together with 220 ng of backbone, 110 ng of insert, 880 U T4 Ligase, and T4 ligase buffer in 44 uL reactions. A total of 4 ligation reactions were run for each library. In addition, a vector-only ligation was set up for

each library with the same conditions, but no insert (to measure the rate of empty vectors self-ligating). The ligation was carried out at 4°C overnight, followed by 65°C for 10 minutes. For library 3, the vector backbone wasn't created *via* direct digestion of the wildtype plasmid; rather, the backbone was amplified *via* PCR while simultaneously adding Esp3I restriction sites as primer overhangs. This backbone was digested with DpnI, then gel extracted. Golden Gate Assembly was conducted in 20 uL reactions with 26 ng vector DNA, 12 ng insert DNA, T4 DNA ligase buffer, 10 U Esp3I, and 400 U T4 DNA Ligase. A vector-only Golden Gate was also tested for this library. Thermal cycling conditions were (37°C for 5 min, 16°C for 5 min) x30 cycles, 55°C for 5 min, and 80°C for 10 min. For all three libraries, the assembled plasmids and empty vector controls were cleaned and concentrated *via* ethanol precipitation with yeast tRNA (ThermoFisher) as a co-precipitate (12 ug yeast tRNA per 300 ng assembled DNA per 840 uL precipitation). The final DNA + tRNA was resuspended in 4 uL ddH2O and nanodropped to confirm ~100% of nucleic acids were retained throughout the precipitations. For 4 transformations per library, 1 uL of this concentrated DNA + tRNA was transformed into 20 uL of ElectroMAX Stbl4 cells according to manufacturer's instructions (ThermoFisher). The transformation outgrowth was incubated for 1 hour at 30°C and 200 rpm. After the outgrowth, all transformations for each library were pooled together, then plated on 8x large petri dishes (15 cm diameter) containing LB agar + kanamycin. Serial dilutions of the transformations were also plated on LB agar + kanamycin to allow colony counting for library coverage calculations. Plates were incubated at 30°C for approximately 24 hours. For each library, all colonies from the large plates were scraped and pooled together in 4 mL 2YT + kanamycin per plate. These cell suspensions were midiprepped (GenElute HP Plasmid Midiprep Kit, Millipore Sigma), and the final pools of library DNA were used for lentivirus production and titering at the University of San Francisco Viracore or Duke University Viracore. When we received the virus, it was divided into single-use aliquots (to avoid repeated freeze-thaw cycles), then stored at -80°C.

3.2.3 High-throughput pooled screen of TdT libraries in HEK293Ts

HEK293Ts were acquired from ATCC (CRL-3216) and cryopreserved in liquid nitrogen upon receipt. To start the high-throughput screen, the cells were thawed in a 37 °C bead bath for approximately 5 minutes, then resuspended in 10 mL pre-warmed DMEM with high glucose and GlutaMAX Supplement (Gibco, 10566024), plus 10% FBS (Sigma, 12306C) and seeded into a cell culture treated T-75 flask (Corning, 430641U). Culturing occurred at 37 °C and in 5% CO2. The initial population of 293Ts was grown long enough to passage once before transducing with lentivirus (5 days). Transduction was carried out in 3x T-150 flasks per library. When the cultures were approximately 30% confluent, the lentiviruses encoding the TdT libraries were transduced at a multiplicity of infection (MOI) of 0.3, with 8 ug/mL polybrene (Fisher Scientific, TR1003G) in the culture media. 24 hours post-transduction, the cells in each T-150 were passaged 1:3 into regular media, in new T-150s (a stringent dilution was avoided, to avoid losing library members). At 48 hours post-transduction, the media was replaced with fresh culture media containing 2 ug/mL puromycin (InvivoGen, ant-pr-1). After approximately 1.5 days of selection, dead cells were removed by refreshing the puromycin-containing media. Cultures were passaged as needed for a total of 6 days in the presence of puromycin. For transfection with Cas9 and an sgRNA designed to generate DSBs downstream of the TdT genes, each library was passaged into 3x T-150 flasks, aiming for 10% seeding density. Approximately 24 hours later, 11.1 ug of a Cas9-expressing plasmid (Addgene #41815), 7.4 ug of the appropriate sgRNA-expressing plasmid (Supplementary Table 7), and 1 ug of an mCherry-expressing plasmid (included for a visual indicator of the transfection's success) were transfected into each flask with FuGENE HD transfection reagent (Promega E2312), according to the manufacturer's instructions, with a 15 minute complexing time. 16 hours after transfecting, the culture media in the flasks was refreshed. 3 days post-transfection, the cultures were trypsinized, resuspended in regular culture media, and a small fraction (4% of the cells) was seeded into a T-25 flask for

Mycoplasma testing (Applied Biological Materials, C214). The remainder of the cells were washed once with phosphate buffered saline (ThermoFisher, 10010023), then frozen at -80°C for downstream DNA extraction and sequencing. All Mycoplasma tests were negative.

3.2.4 Cell culture methods for individual TdT assays and CHYRON experiments

HEK293Ts were procured directly from ATCC (CRL-3216) but were not otherwise authenticated. All cell culture was conducted in DMEM with high glucose plus GlutaMAX (Gibco, 10566024) and 10% FBS (Sigma, 12306C). Cultures were maintained at 37 °C and in 5% CO2.

For individual TdT variant assays, HEK293Ts were thawed from our liquid nitrogen collection, using a bead bath at 37 °C for approximately 5 minutes. Cells were gently resuspended in pre-warmed culture media and seeded into cell culture treated dishes. 1-2 days later, cultures were split into 6-well dishes (Falcon, 353046) with a seeding density of approximately 5%. The following day, hp-minipreps of Cas9 (750 ng), the HEK3-targeting sgRNA (500 ng), and the TdT variant of interest (750 ng) were transiently transfected using FuGENE HD Transfection Reagent (Promega E2312) according to manufacturer instructions. 20-minute complexing times were found to be appropriate for these constructs. 3 days post-transfection, the cells were collected by aspirating the media out of the cultures, spraying cells off the plates with cold PBS, pelleting the cells, removing the supernatant, and placing cell pellets at -80°C for later processing.

For CHYRON experiments, we used our previously reported $CHYRON_{20i}$ cell line[22], which possesses a homing guide RNA (hgRNA) gene with a 20 nt spacer integrated at the AAVS1 locus and insulator flanks. Cas9-T2A-TdT transgenes with the TdT variants of interest were either stably integrated into the genome with Super PiggyBAC Transposase (System Biosciences Innovation, PB210PA-1) (for continuous propagation of the CHYRON locus as in Figure 3.4b-c) or transiently transfected at 4-day intervals (for signal recording as in Figure 3.4d-

g). In both cases, FuGENE HD Transfection Reagent with a 20-minute complexing time was used to introduce the transgenes into the cells (Promega E2312). For stable cell lines, samples were passaged into 6-well dishes 1 day prior to transfection, with a seeding density of approximately 20%. Transfection was performed with a 5:1 cargo: transposase ratio. 2 days later, samples were passaged 1:10 into media containing 10 ug/mL blasticin (Thermo Scientific, J672168EQ), to select for successfully transformed cells. An untransfected control was also cultured in 10 mg/mL blasticin, to confirm the selection killed ~100% of non-resistant cells before the first timepoint was collected. Blasticidin selection was maintained throughout the entire timecourse. Cultures were passaged approximately once every 3 days, and cell pellets containing at least 300,000 cells (corresponding to 10% confluency in a 6-well dish) were periodically sampled for NGS. For transiently transfected samples (signal recording experiment), the CHYRON$_{20i}$ cells were passaged into 24-well dishes 1 day prior to transfecting. FuGENE HD transfection reagent (Promega E2312) was used as specified in the user manual to transfect 400 ng of the hypoxia and/or Wnt recording cassettes into the 24-well cultures, with a 20 min complexing time. 16 hours after transfecting, the culture media was aspirated off, and replaced with either regular culture media or inducer-containing media. Recombinant human Wnt-3a protein (R&D Systems, 5036-WN-010) was used to induce Wnt recording at a final concentration of 250 ng/mL in the culture media, unless stated otherwise. The HIF-hydroxylase inhibitor, DMOG (Millipore Sigma, 40009150MG), was used to induce hypoxia recording at a final concentration of 1 mM in the culture media, unless stated otherwise. When real hypoxia was used, as in Figure 4g, a 3% O2, 5% CO2, balance N2 environment was created by placing the culture dishes in a Modular Incubator Chamber (Embrient, Inc., formerly Billups-Rothenberg, Inc., MIC-101) and flushing the headspace with the desired gas mixture for 3 minutes. A Dual Flow Meter (Embrient, Inc., formerly Billups-Rothenberg, Inc., DFM-3002) was used to achieve the appropriate influx from two different compressed gas cylinders—one cylinder contained 5% CO2 / balance O2 (Airgas, #Z02OX9512000033), and the other contained 5% CO2 / balance

N2 (Airgas, #X02NI95C2003071). Cultures were re-transfected with the appropriate recording cassettes every 4 days, and re-induced again approximately 16 hours after transfecting.

3.2.5 Illumina library preparation for the high-throughput screen

For each of the 3 site saturation mutagenesis libraries, all the cell mass from the end of the *in vivo* assay was subjected to genomic DNA (gDNA) extraction with a QIAamp DNA Mini kit (Qiagen, 51304), then combined into one pool per library. A "1st round PCR" was used to amplify the sequences of interest from the gDNA and add partial Illumina adapters. Optimal conditions for the 1st round PCR were explored by testing 2 alterative sets of primers for each library, 4 different annealing temperatures, 4 different extension times, and 4 different cycles of amplification (20, 25, 30, and 35 cycles) using 600 ng of template gDNA per 50 uL reaction with Phusion Hot Start Flex polymerase (New England Biolabs, M0535). Next, using the optimized PCR conditions, we determined the maximal amount of gDNA template that could be tolerated in the 1st round PCR and still yield robust amplification (defined as producing visible banding on an agarose gel after 25 cycles). A high mass of genomic DNA per PCR was desirable because it is thought to reduce the risk of PCR jackpotting[28], and it would allow us to sample gDNA from a massive number of cells, improving our chances of capturing at least 250 insertions from each TdT genotype. We found 2, 4, and 4 ug per PCR to be appropriate for libraries 1, 2, and 3, respectively. Having settled on these conditions, we set up the 1st round PCR at scale for each library, using 3 sets of multiplexing primers per library (barcodes on both F and R primers). 25 replicate PCRs were run with each pair of multiplexing primers (75 total PCRs per library). Using multiple sets of uniquely barcoded primers allowed us to both identify instances of template switching (*via* nonsensical combinations of F and R barcodes on template-switched) and assess the trustworthiness of our data by comparing results across the 3 differently barcoded technical replicates (Supplementary Figure 1). Phusion Hot Start Flex polymerase (New England Biolabs, M0535) was used for the 1st round PCR, for 20 cycles of amplification with 62°C, 63°C, and

66°C annealing temperatures for libraries 1, 2, and 3, respectively, and an extension time of 40 seconds. The 1st round PCR products were pooled together for each library, then cleaned with a 0.9:1 ratio of AMPure XP beads:PCR mix (Beckman Coulter, A63881). Then, a "2nd round PCR" was performed to add complete Illumina adaptors. The complete volume of the AMPure-cleaned 1st round PCR products for each library were added to a 2nd round PCR master mix, which was then divided across 10 replicate reactions for each library. Again, the use of many replicate PCRs was intended to curb PCR jackpotting. Phusion Hot Start Flex amplified the 2nd round PCRs in 20 rounds of thermal cycling with a 63°C annealing temperature and 40 second extension time. The final PCR products were pooled (one pool for each library), cleaned again with 0.9:1 AMPure XP beads: PCR mix, and sequenced with an Illumina NovaSeq6000 using the Paired End 250 protocol at the UC Irvine Genomics, Research, & Technology Hub.

3.2.6 *In vitro* expression and purification of TdTs.

T7 express *E. coli* (NEB) was transformed with the final pET28a constructs. Transformation outgrowth was plated on kanamycin and incubated overnight at 30°C. Colonies were selected and grown overnight in 5 mL LB media with 25 ug/mL kanamycin. The next day, these overnight cultures were diluted 400x in 60 mL of fresh LB media and 25 ug/mL kanamycin. The cultures were grown to an OD of 0.4-0.8 (~3 hours) at 37°C, induced with 1 mM of IPTG, and incubated for 16 hours at 15°C. After 16 hours, the cultures were centrifuged at 4000xg for 20 minutes, and the supernatant was discarded.

To purify TdT, TALON Metal Affinity Resin (Takara) protocols were used, and all steps were performed at 4°C on ice. To lyse the harvested cells, pellets were resuspended in xTractor buffer (20mL per 1g pellet) with 1X lysozyme (200uL of 100X lysozyme solution per 1g pellet) and DNase I solution (40uL of 5units/uL per 1g pellet). After incubating on a shaking platform for 20 minutes at 175 RPM, the tubes were centrifuged at 12,000xg for 20 minutes. The

supernatant (soluble fraction) was applied to 100 uL of equilibrated TALON Resin and incubated on a shaking platform for 20 minutes to allow the His-tagged TdT to bind the resin. Thereafter, the resin was pelleted by centrifugation (700 xg for 5 minutes at 4°C) and the supernatant was carefully removed. The resin was washed twice by incubating it with 1mL equilibration buffer for 10 minutes (Takara), centrifuging the resin (700xg for 5 minutes) and discarding the supernatant. After washing, the resin was resuspended in 100 uL of equilibration buffer, briefly vortexed, and transferred to a 2 mL gravity column (Takara). The end-cap of the column was removed to drain the buffer and 500 uL of wash buffer was added to wash the resin once more. 500 uL of Elution Buffer (Takara) was then added to the column to elute the TdT into a clean 1.5 mL tube. The protein concentration was measured using a Nanodrop (absorbance at 280nm). TdT expression and purity was verified and analyzed via SDS-PAGE. The elution was buffer-exchanged and concentrated using 10K Amicon-Ultra 0.5 centrifugal filters (10,000 MWCO) into TdT storage buffer (200 mM KH2PO4 and 100 mM NaCl at pH 6.5). The fractions were aliquoted and stored at –80°C.

3.2.7 *In vitro* TdT extension assays

All TdT extension reactions were set up on ice and contained final concentrations of 0.1 uM of common sequence 1 primer (CS1_5N, Supplementary Table 2), which has 5 degenerate bases at the 3’ end of the primer. 400 uM of dNTP mix (100 uM of each), 1X NEB TdT reaction buffer, purified TdT enzyme (varying amounts depending on the variant), and nuclease-free water were added to 50 uL. The reactions were performed for 1 hour at 37°C. Reactions that were analyzed *via* urea-PAGE gels were performed with FAM-labelled primer; reactions performed for NGS analysis used a non-fluorescent CS1 primer.

For each variant, a range of TdT concentrations were tested to titrate TdT such that the approximate average extension lengths after 1h at 37°C were in the appropriate length range for

next-generation sequencing. We aimed for an average extension length below ~60nt. The final concentrations used for extension reactions that were sequenced are shown in the table below.

TdT variant	TdT yield following purification (mg/ml)	Final TdT stock concentration following buffer exchange (mg/ml)	Final TdT concentration in extension reaction for NGS (ug TdT/50uL reaction)
Wildtype TdT (pCKC755)	0.26	0.44	0.2
Unbiased1 (pCKC768)	0.14	0.31	0.2
Unbiased2 (pCKC769)	0.18	0.41	0.2
Unbiased3 (pCKC770)	0.31	0.47	0.3
A/T-biased1 (pCKC771)	0.21	0.25	0.1
A/T-biased2 (pCKC772)	0.27	0.38	0.1
G/C-biased (pCKC773)	0.27	0.31	0.3
Wildtype-like (pCKC774)	0.25	0.40	0.2

The extension products were analyzed by urea-PAGE. 8uL of the completed extension reaction was mixed with 12 uL of BioRad 2x TBE urea sample buffer and boiled for 10 min. The samples were loaded into a 10% polyacrylamide TBE urea gel (bioRad 4566036) and run at 200 V for 40 min. The gels were imaged on a Sapphire Biomolecular Imager using 100 um pixel size and λex = 488 nm and λem = 520 nm. The imaging gain was adjusted for each experiment to avoid saturation.

3.2.8 Next-Generation Sequencing of *in vitro*-extended samples

To prepare the extension reactions for NGS, common sequence 2 (CS2) primer was first ligated to the 3' end of the TdT-synthesized polynucleotides. 2uL of the extension reaction product was used in a ligation reaction containing final concentrations of 1 uM CS2 primer, 10 units of T4 RNA ligase (NEB), 1X T4 RNA ligase reaction buffer (NEB), 10 uL of PEG 8000 (NEB), 1 mM of

ATP, and nuclease-free water up to 20 uL reaction volume. The ligation reaction was carried out at 25°C for 2 hours.

PCR was performed to prepare the ligation products as double-stranded, barcoded samples ready for sequencing according to previous protocols[17]. Primer sets from the Access Array Barcode Library for Illumina Sequencers (Fluidigm) were used for barcoding. The Fluidigm primer sets contain the complement of CS1 (the defined sequence of CS1_5N) as the binding target of the forward PCR primer and the complement of CS2 as the reverse primer. The Fluidigm primers also contain a unique barcode in the reverse primer and Illumina adapters in the forward and reverse primers for sequencing. The PCRs consisted of 2 uL of the 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 uL 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. PCR samples were sent for sequencing at Rush University's Genomics and Microbiome Core Facility. Quality control (QC) and Illumina MiniSeq protocols were the same as described previously[17].

3.2.9 Computational analysis of NGS data

All analyses were conducted with custom Python scripts, which will be made available at github.com/liusynevolab/TdT_high-throughput_screen.

For the high-throughput screen, the amplicons that were sequenced were ~1 kb long (sufficiently long to cover the NNK codons in the TdT gene near the 5' end of the amplicons, and the Cas9 target site near the 3' end), so the Illumina reads (paired end 250) did not cover the full length of the molecules. Thus, forward and reverse reads for each amplicon were simply concatenated, such that the NNK sequences were stored in the same string as the TdT-

mediated insertion at the Cas9 target site. In other words, nothing of importance resided in the middle of the ~1 kb amplicons, and there was no sequence overlap between the forward and reverse reads, so there was no need to align them.

After concatenation, reads of incorrect amplicons (*e.g.*, primer dimers or non-specific amplicons) were removed. Extremely short amplicons, such as primer dimers, were identifiable without performing a formal sequence alignment because they were marked with long repetitive stretches of "G" in their reads—this is because Illumina uses sequencing by synthesis with a fluorescent readout to indicate which base was incorporated in each round of synthesis. Short molecules generate no fluorescence in the late rounds of sequencing by synthesis, and lack of fluorescence was miscalled as guanine by the Novaseq basecalling software. Thus, any reads with homopolymeric stretches longer than 15 bp were automatically removed. 15 bp is the maximum insertion size we attribute to one round of TdT insertion mutagenesis (anything larger tends to be due to homologous recombination at the Cas9 cut site), so this cutoff would tolerate the hypothetical case of an extremely biased TdT inserting 15 bp of a single nucleotide.

Once the non-specific short reads were filtered out, the remaining reads were demultiplexed according to the barcodes on the Illumina sequencing primers. Each library was labeled with 3 different pairs of multiplexing barcodes, so a total of 9 barcode pairs were expected across the 3 libraries. Any reads that had unexpected barcode pairs were assumed to be the product of template-switching during library preparation and were discarded.

The demultiplexed reads were aligned to an appropriate reference sequence for each library, using Mapp as in Perli et al[29]. Alignments are represented by Mapp as strings containing "M" for match, "X" for mismatch, "D" for deletion, and "I" for insertion at each base in the read. These alignments were used for another quality filtering step, wherein reads that had >50 mismatches (>10% of bases in the read were mismatches), >250 deletions (>half of the amplicon was deleted), or >15 contiguous insertions (explained above) were removed.

The DXMI characters were also used to extract NNK sequences and insertion sequences from the aligned reads. Codons that were mutagenized in each library were designated with “NNN” in the reference sequence, so the alignment always indicated three mismatches (“XXX”) when aligning reads with the randomized codons. Thus, the analysis script searched for the expected alignment at the sites of saturation mutagenesis, and checked for 2 perfectly aligned bases (“MM”) on either side of the mutated residues (“MMXXXMM”). Reads that passed all the filters to this point, and contained the expected alignment pattern, were deemed reliable enough to extract their NNKs. Similarly, the DXMI characters were used for extracting TdT-mediated insertion sequences and NHEJ-generated deletions from the Cas9 cut site in each read. A small window around the expected location of the Cas9 cut site was checked for the presence of an “I” or “D” in the mapp alignment. If an “I” was present, the adjacent alignment characters were scanned for additional, contiguous “I’s”, so the full length of each insertion was captured. Once the entire string of “I’s” was found, the script ascertained that the read had perfectly aligned bases on either side of the insertion (“MM I_n MM”), and if so, it extracted the insertion sequence. Any Cas9 target site that harbored one or more “D(s)” in the mapp alignment was counted as a deletion. The sequences at each Cas9 cut site were sorted according to which TdT mutant gene (NNK codons) was linked on the same read. Ultimately, each TdT variant had a collection of hundreds to thousands of editing outcomes associated with it.

Having binned all insertions and tallied all deletions for each TdT variant, the script quantified the following metrics for each TdT variant: total number of extracted insertion sequences, average length of insertions, average A/T bias among all inserted nucleotides, average A/T bias for each insertion length, number of insertion sequences of each length, “activity fraction” (*i.e.*, number reads with an insertion / number of reads with any indel), and number of unique insertion sequences for each TdT variant. For each library, a comprehensive dataframe was created, reporting all of these metrics for each TdT variant. Excel was used to

manually inspect the dataframe, sorted by overall A/T bias of each TdT variant, and select variants of interest for follow-up experiments. Matplotlib was used to create histograms of activity fractions and scatterplots of A/T bias vs. average insertion length for each TdT variant.

For individual assays of TdT variants and CHYRON experiments, samples were demultiplexed *via* sample-specific barcodes on the forward and reverse reads. Insertions were extracted by checking each reach for the expected 15-bp sequence upstream of the Cas9 target site (within a user-defined Hamming distance), then scanning for the 15-bp sequence downstream of the Cas9 target site (within a user-defined Hamming distance), and finally grabbing the intervening sequence. If this initial alignment failed to locate the Cas9 target site (*e.g.*, because the locus was partially deleted, a normal editing outcome for Cas9 cutting), it searched for an alignment with 15-bp sequences further out from the cut site. The distance between the upstream and downstream aligned sequences was used to calculate the size of deletions. The lengths of all insertions and deletions were tallied, to create a histogram of indel sizes. Compositions of the extracted insertions (average insertion length, overall A/T content across every inserted nucleotide, and average usage of all 4 natural nucleotides for each insertion length) were reported.

3.2.10 Mathematical considerations for site saturation mutagenesis libraries

For early stages of the library screen, the goal was to represent every possible randomized codon sequence throughout cloning, lentiviral production, and HEK293T transduction. Each library was cloned with 3 randomized codons, which were encoded by NNKs (N = any nucleotide, K = G or T). Therefore, the number of possible DNA sequences in each library was $(4*4*2 \text{ possible NNK sequences})^{(3 \text{ NNK codons})} = 32{,}768$ sequences. As previously described, 12x library coverage establishes a 99% likelihood of every possible DNA sequence being captured[30], so we aimed to hit a minimum of 12x coverage during cloning, but we overshot this goal as much as practically possible. In the end, each library was covered 70.7-80x. The hp-

minipreps encoding the libraries were assumed to be more than sufficient to cover the libraries, since a single nanogram of our 10.3 kb library plasmid should contain approximately 9.5E+07 molecules (3000x library coverage). After the libraries were packaged into lentivirus, the titers were 1.6E+08, 8.5E+08, and 8.4E+08 infectious particles / mL, for libraries 1, 2, and 3, respectively. At least 100 uL of each virus was provided by the UCSF ViraCore and Duke University ViraCore, constituting approximately 500x library coverage. During transduction, we aimed for a minimum of 12x library coverage, to practically guarantee every library member would be present, and we exceeded the 12x coverage goal as much as was feasible. Ultimately, 3x T-150 flasks at ~30% confluency were transduced at a MOI of 0.3 for each library. Assuming 3,300 cells / mm^2 in a confluent culture dish (empirically approximated using a hemacytometer), each transduction produced nearly 4.5E+06 infected cells (137x library coverage).

After selecting for transduced cells, our goal was to not only cover every library member, but to cover every member enough times to produce a useful degree of confidence in the average behavior of each library member. In other words, we decided how many insertion sequences needed to be attained from each library member. Considering a simple scenario where each TdT variant inserts only a single nucleotide at each Cas9-mediated double strand break, we could model each insertion as a Bernoulli process (*i.e.*, independent experiments with a binomial outcome that occurs with a probability of p). In this model, we consider the binomial outcome to be an A/T insertion (a “success”) or a G/C insertion (a “failure”), and the probability of an A/T insertion equal to the A/T bias of the library member. Therefore, the probability of having exactly k insertions comprised of A or T in n trials is equal to:

(1) $$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Where Pr is the probability that the random variable X is equal to k successes (number of A/T insertions) in n trials for a library member with an A/T bias of p. This probability function was used to calculate the fiducial limits of p given an observed k / n, with the Clopper-Pearson

method[31]. We assessed the 95% confidence intervals for different A/T bias values that we hoped to encounter in the library. We found 250 trials per library member to strike the right balance between being practically feasible and providing an acceptable degree of certainty in the data. The 95% confidence intervals for the true A/T biases, given an observed % A/T are shown below:

Observed % A/T in TdT-mediated insertions (*i.e.*, k successes / n trials)	**Actual value of p (*i.e.*, A/T bias), 95% confidence intervals**
10%	5-15%
50%	44-56%
90%	85-95%

Although most TdT variants were expected to make longer than 1 bp insertions on average (wildtype insertions are ~3 bp), TdT's insertion compositions are known to be affected by the previously inserted nucleotides[32]. Therefore, we maintained a requirement of 250 total insertions per variant (instead of 250 inserted nucleotides) because each insertion is independent of every other insertion. To obtain approximately 250 insertions per library member, we accounted for the transfection efficiency of Cas9 and its sgRNA (we usually see ~80% overall editing efficiency), and the frequency with which TdT usually makes an insertion at the Cas9-generated DSB (approximately 75% of DSBs are repaired with an insertion). Thus, the equation to determine the number of necessary cells to transfect with Cas9 and its sgRNA is:

(2) $(250\ insertions\ per\ variant) * (32{,}768\ variants\ per\ library) * \left(\frac{100\ cells}{80\ edited\ cels}\right) * \left(\frac{100\ edited\ cells}{75\ insertions}\right) = 1.4E+07\ cells\ per\ library$

At this stage, we intentionally surpassed the number of cells in equation 2 as much as was experimentally feasible, to safeguard library coverage. For each library, 3x T-150 flasks, each approximately 25% confluent, were transfected. This amounts to approximately 1.25E+07 cells

at the time of transfection. We assumed Cas9 would be expressed and start cleaving its target locus at 24 hours post-transfection, at which point the cultures would have 2.5E+07 cells (nearly 2x coverage of the number of cells calculated in equation 2). The cultures were incubated for a total of 72 hours after transfection, which is long enough to allow Cas9 cutting, TdT inserting, and DNA repair, but short enough to prevent excessive outgrowth. Each cell should have divided only 1-2 times after acquiring a TdT-mediated insertion. All cells in a given library were pooled together upon collection.

The genomic DNA (gDNA) from the complete pool of HEK293Ts was extracted and pooled together. For Illumina prep PCRs, we used the maximum amount of gDNA as a template for each PCR, and as many PCRs as was feasible. Ultimately, we used 150 ug, 300 ug, and 300 ug genomic DNA as template for libraries 1, 2, and 3, respectively. Assuming 8 pg gDNA per HEK293T cell, this corresponds to 1.9E+07, 3.8E+07, and 3.8E+07 cells' gDNA in the PCRs for Illumina library prep. The number of PCR amplification cycles was minimized during library prep, to reduce the production of duplicate sequences (we found 20 cycles to be sufficient for both rounds of PCR, see Methods). Final Illumina libraries were sequenced at 22x coverage of 250 insertions per sample, for a total of 3.1E+08 reads. Because our Illumina libraries had a larger molecular weight (~1 kb) than typical Illumina libraries (<500 bp), our target amplicons clustered somewhat inefficiently on the NovaSeq flow cell. After filtering out primer dimers and low-quality reads, we retained 1.0E+08 Illumina reads, and we extracted averages of 175, 215, and 367 insertion sequences per TdT variant (variants defined by DNA sequences) for libraries 1, 2, and 3. When synonymous TdT variants were binned together, we had averages of 1131, 1402, and 2302 insertions per TdT variant (variants defined by amino acid sequences).

Although the total number of insertions per TdT variant was satisfactory (exceeding 250) for most variants (defined by amino acid sequences), we also wanted to assess what fraction of the insertions were likely to originate from independent TdT-mediated insertion events (rather than being the product of PCR duplication or Cas9 target site duplication during HEK293T

outgrowth). For this, we looked to the unbiased TdTs in each library (overall A/T bias within the range 49-51%), since they are least likely to generate identical insertion sequences in independent trials. Also, there is an inherent link between the length of an insertion mutation and the probability that it will coincide with another insertion from an independent event. To determine what length insertions could be expected to have 100% unique identities if they all arose from independent events, we used a form of the Birthday Problem that defines the probability of no sequences being identical:

(3) $$P = 1 \times \frac{4^L - 1}{4^L} \times \frac{4^L - 2}{4^L} \times \ldots \times \frac{4^L - (n-1)}{4^L}$$

Where P is the probability of all sequences being unique, L is the length of the insertion sequences, and n is the number of insertion sequences being compared to each other. Using equation 3, we found that 6-bp insertions are 95% likely to be entirely unique if there are 20 insertion sequences being compared. Likewise, 7- and 8-bp insertions are 95% likely to be unique if there are 40 and 80 insertion sequences being compared, respectively. Insertions were first binned according to which TdT variant made the insertions, then the number of insertions in each bin was counted. If a variant had >20, >40, or >80 insertions of 6, 7, or 8 bp, respectively, the insertions in the problematic size bin could not be used to assess the fraction of unique insertions (because it is statistically likely that independent insertions would be identical by chance). For example, if a TdT variant had 30x 6-bp insertions (more than the cutoff of 20x 6-bp insertions), but only 25x 7-bp and 15x 8-bp insertions, the 6-bp insertions couldn't be used to determine the fraction of independent insertions, but the 7-bp and 8-bp insertions could be used. Every bin of 6-, 7-, and 8-bp insertions corresponding to unbiased TdTs, with fewer than 20, 40, and 80 insertions in the bins, were checked for their fraction of unique insertions in the bins. The average fraction of unique insertions was assumed to extend to the entire library (*i.e.*, if 50% of the insertions were found to be unique for the unbiased TdTs, we

assumed this meant 50% of all insertions in the screen originated from independent insertion events).

3.3 Results

3.3.1 Three site saturation mutagenesis libraries were assayed in high throughput, uncovering TdT mutants with altered nucleotide preferences

While the massively parallel screen presented here has the potential to facilitate a variety of attractive TdT engineering endeavors, we focused on one that is pertinent to *in vivo* DNA recording. Specifically, we wanted to upgrade our existing DNA recorder, CHYRON[22], by creating a collection of TdT variants with altered nucleotide insertion preferences. Wildtype TdT has a natural, moderate bias for incorporating G's and C's[2] (*i.e.*, we observe its A/T bias is 35-40% when used in CHYRON), but an unbiased TdT would be ideal for cell lineage recording, and strongly biased TdTs would enable multiplexed recording of molecular events. To manipulate the bias of TdT, three site saturation mutagenesis libraries were devised, each targeting three residues that were hypothesized to affect nucleotide preferences (Figure 3.1a-b). Two of the libraries targeted residues in a 16 amino acid lariat-like loop that is unique to TdT and known to be responsible for its template-independent activity (Loop1)[2]. We suspected Loop1 may influence dNTP selection because it adopts various conformations in the available crystal structures of TdT, suggesting it may play a dynamic role in polymerization. In one library (Library 1), we randomized three codons at the base of Loop1 (D399, H400, F401) because they could feasibly alter the range of motion of the loop. In another library (Library 2), residues in Loop1 that we hypothesized could move close enough to the incoming dNTP to interact with it directly (D396, L398, D399) were targeted. Aside from Loop1, we also mutagenized residues on an alpha helix, which reside within 6 Å of the bound nucleotide and point toward the nucleobase

(R454, E457, R461)[2], and are thus suspected to influence nucleotide preferences *via* interactions with the base (Library 3).

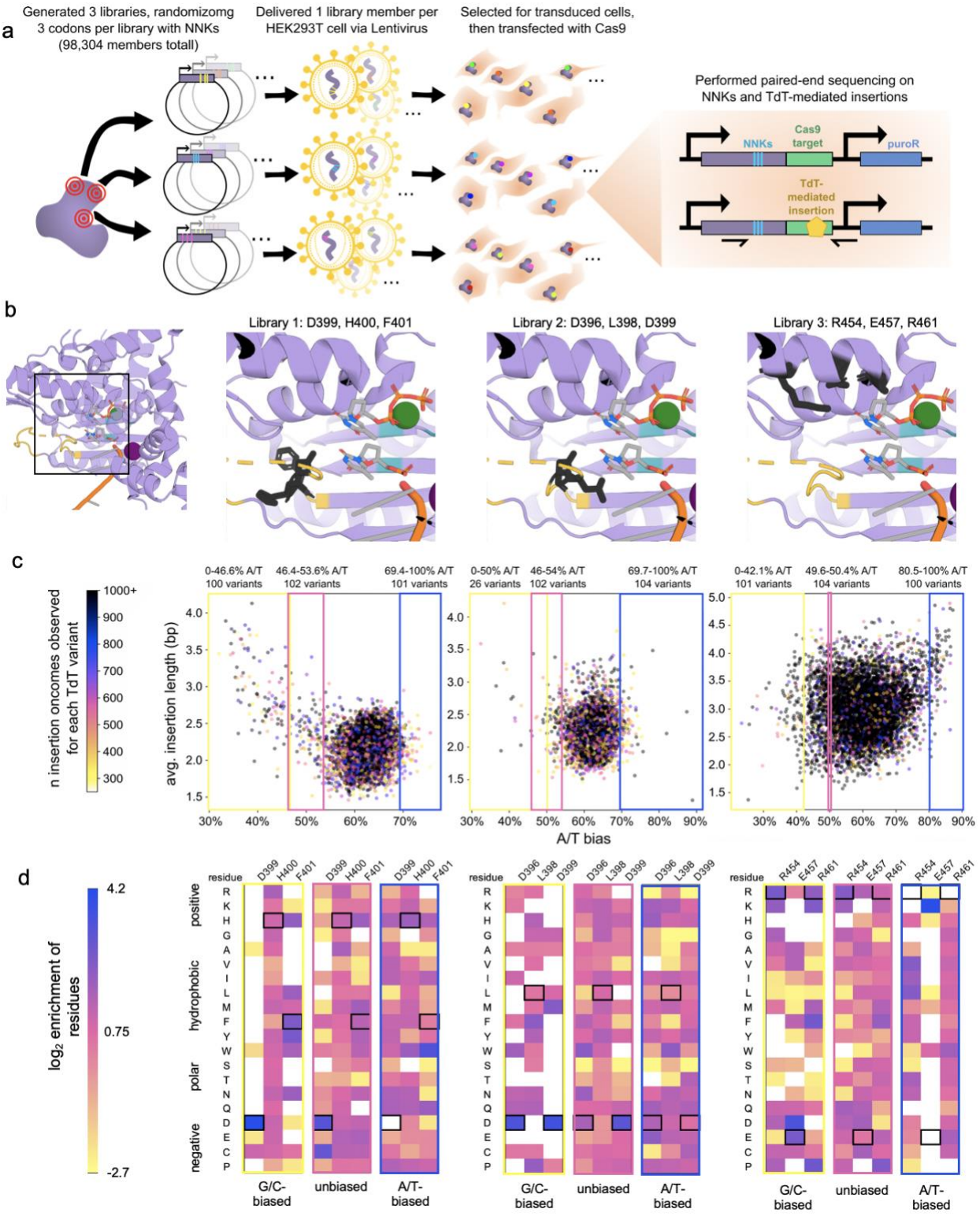


Figure 3.1. A massively parallel assay on 3 site saturation mutagenesis libraries of TdT yielded variants with altered nucleotide preferences. **a,** Workflow for creating the saturation mutagenesis libraries encoded in lentivirus, and assaying the library members in HEK293Ts. Each TdT variant is assayed hundreds to thousands of times, producing statistically robust readouts on the properties of each variant. **b,** The mutagenized residues in each library are shown in crystal structures of TdT bound to a single-stranded DNA primer (single helix colored with red, orange, and gray) and an incoming ddTTP (red, blue, and gray thymidine with red and orange triphosphate moiety). Mutagenized codons are colored black and represented with stick structures. Loop1 is highlighted in yellow. The catalytic aspartates are indicated in teal. Green and plum spheres are divalent magnesium and sodium, respectively. **c,** Diverse combinations of A/T biases and average insertion lengths are present among library members. Single points represent individual TdT variants. Color-coding reflects the number of insertion sequences analyzed for each variant. Yellow, pink, and blue boxes demarcate bins of the ~100 most G/C-biased, unbiased, and A/T-biased TdTs in each library. **d,** For the bias bins indicated in c, the composition of residues at the randomized loci were assessed. Heatmaps show the enrichment for each residue relative to the full library of active variants. White cells indicate residues that were absent in the bin (infinitely de-enriched).

Randomizing 3 residues each in 3 separate libraries produces a total of 27,783 distinct protein sequences, corresponding to 98,304 mutant TdT DNA sequences. Assaying libraries of this magnitude necessitated the development of a massively parallel screen for beneficial mutations (Figure 3.1a). A pooled approach was possible because the phenotype we aimed to achieve (altered nucleotide insertion preferences) could be read out *via* next-generation sequencing (NGS), and the target site for TdT-mediated insertions could be user-specified (TdT readily produces insertions at Cas9-generated double strand breaks (DSBs)[4,22]). The latter consideration was critical, as it allowed the phenotype of each TdT mutant to be linked to its genotype by placing both the randomized loci in the TdT gene, and the TdT-mediated insertion sequences, on the same molecule for NGS. In addition to enabling the phenotype-genotype linkage, TdT's amenability to editing genomic DSBs allowed us to assay its activity in live mammalian cells. The use of mammalian cells was desirable because it allowed lentivirus to be the vector for library delivery, which is a previously validated approach to screening pooled libraries by delivering precisely 1 library member into each cell (only ~5% of cells will receive more than one library member)[33]. With this method, each cell encapsulates one mutant TdT gene and the corresponding reaction of interest. This physical separation between library members ensures the readout of inserted nucleotides in each cell can be reliably attributed to

the correct TdT variant. Also, scaling the assay to achieve the desired number of instances of each library member is straightforward when using mammalian cells. The size of the cell culture and the corresponding amount of lentivirus are simply scaled up enough to ensure each library member is tested in many independent trials.

Throughout library synthesis and the cell-based assay, library diversity was maintained by erring on the side of excessively large sampling depth. For site saturation mutagenesis, we PCR-amplified the TdT gene using primers that contained the degenerate bases NNK at the codons we wished to randomize. N represents any nucleotide and K represents G or T. The last nucleotide in the randomized codons is designated with K rather than N because this allows all 20 natural amino acids to be included, while reducing library size and avoiding stop codons[30]. To ensure every possible sequence was included in each library, the libraries were substantially over-sampled during cloning. 12x coverage of each library would establish a 99% likelihood of every possible DNA sequence being captured[30], and we exceeded this rule of thumb with 70.7-80x coverage of each library (Supplementary Table 1). Also, to confirm the observed library size was an accurate reflection of viable library members, we measured the rate of empty vectors self-ligating (without a TdT gene being inserted) by performing all library cloning steps on vector-only controls of each library. This projected a low rate (0.52-2.6%) of background due to vector reclosures. The libraries were packaged into lentivirus, which was then transduced into a large population of HEK293T cells. Two important considerations factored into the transduction step—enough cells and viral particles were deployed to promote sampling of every library member (again, >12x coverage), and the ratio of cells to viral particles was tuned to achieve only one lentivirus on average infecting each cell (we used an MOI of 0.3). The latter consideration ensured only one TdT variant is expressed per cell. Puromycin was used to select for infected cells. Lastly, Cas9 and a sgRNA targeting a locus immediately downstream of the TdT gene were transfected into the cells (Figure 3.1a). This produced the necessary conditions for every library member to be assayed in many independent instances, simultaneously. During

Cas9 transfection, the objective was to transform enough cells to not only fully cover the libraries, but to test each library member enough times to produce a reasonable degree of confidence in the resulting data for every member. We used a binomial distribution to approximate the number of insertion sequences needed to confidently interpret the A/T biases of each TdT variant (see Methods). Ultimately, we aimed to assay each TdT variant at least 250 times. Each cell expressed one TdT variant, which had an opportunity to make an insertion mutation at the DSB produced by Cas9. Paired-end Illumina sequencing was used to interrogate the NNK sequences in the TdT gene and the Cas9 target sites from the entire pool of transduced cells.

The Illumina library preparation was designed to maximize data yield and quality, and to allow post-hoc assessment of the trustworthiness of the data. Because the NNKs were near the middle of the TdT gene, and the Cas9 target site was downstream of TdT's transcription termination signal, the sizes of the amplicons for paired-end sequencing were larger than typical Illumina libraries (1141 – 1174 bp). We anticipated these large molecules may not efficiently participate in bridge amplification on the Illumina flow cell, so we strived to attain as many usable reads as possible despite the inherent challenges. First, we minimized template switching during Illumina library prep, since template-switched amplicons could provide incorrect linkage between unrelated TdT variants and insertion mutations at the Cas9 target sites, and thus needed to be discarded. Our locus of interest was PCR-amplified during Illumina prep, so we optimized the PCR conditions to achieve robust amplification (there is less opportunity for template switching if each DNA molecule is efficiently synthesized in each round of PCR) (see Methods). We used unique multiplexing barcodes on the forward and reverse primers for each library, so we could ultimately detect the frequency of template-switching during by calculating the occurrence of mismatched barcode pairs and throw out the offending reads. In the final sequencing data, only 0.007% of reads had barcode pairs indicative of template switching; however, we acknowledge our approach could only detect template switching in the later stages

of NGS prep and on the sequencing flow cell. Another consideration for data quality is to avoid jackpotting during PCR. Taking inspiration from others who aimed to Illumina-sequence large, diverse libraries without skewing the representation of library members[34], each PCR was set up with the maximum tolerable amount of template DNA (experimentally determined to be 2, 4, and 4 ug gDNA per reaction for libraries 1, 2, and 3, respectively), and each library was amplified in many separate, replicate PCRs (n=75) to dilute the impact of any single jackpotting event. Furthermore, three different sets of multiplexing primers were used in the replicate PCRs for each library, thereby dividing the PCRs into three discernable pools that could be analyzed as technical replicates (Supplementary Figure 3.1). The properties of each TdT variant could be compared among the technical replicates, and jackpotting is apparent as disagreement among the replicates. When selecting TdT variants for downstream applications, cross-referencing the outputs among the 3 technical replicates allowed false positives due to jackpotting to be easily detected. In the end, the TdT variants had 1131 - 2302 insertion sequences on average extracted from the NGS data ("variant" = protein sequence; synonymous NNKs were binned together to improve the statistical power of the data), and we conservatively estimate 39-50% of the insertions to have originated from independent editing events (as opposed to duplicate sequences arising from HEK293Ts growing after acquiring an edit, or PCR duplications) (see Methods).

The Illumina data revealed that most TdT variants in the pooled assay were both sampled enough times to provide reasonably confident data and maintained high enzymatic activity. As previously mentioned, we chose 250 insertion sequences as the minimum cutoff for trustworthy data. We filtered out all variants that had fewer than 694 reads (this is the minimum number of reads for an active variant with typical rates of editing: 250 insertions / 80% of edited loci bearing an insertion / 80% of loci bearing an edit), so we could approximate the total number of variants that were sequenced deeply enough to have a chance of being selected as appealing mutants. 7204, 7248, and 8728 variants passed this filter in libraries 1, 2, and 3,

respectively, out of a maximum 9261 possible for each library. We also filtered out inactive TdT variants because they aren't useful for practical applications, and they could produce misleading results; that is, small insertions would still be expected at the corresponding Cas9 target sites, but they would merely be indels from endogenous non-homologous end joining (NHEJ). Therefore, the compositions of these insertions would be irrelevant to the activity of TdT. To determine the threshold between active and inactive variants, we first created a metric for activity called the "activity fraction." It is the frequency with which a TdT variant installs an insertion at a DSB. Since DSBs can only be empirically detected if they lead to a mutation, we defined the activity fraction as (reads containing an insertion) / (reads containing any indel). Co-expressing Cas9 with a catalytically dead TdT in HEK293Ts yielded an activity fraction of 0.27, whereas wildtype TdT produces an activity fraction of 0.83 (Supplementary Figure 3.2). This indicates the activity fraction is an effective measure for distinguishing between inactive and active TdTs. A histogram of all TdT variants' activity fractions exhibits a bimodal distribution (Supplementary Figure 3.2), where apparently active variants cluster in the range of 0.7-1, and inactive variants have activity fraction in the range 0.2-0.7. Thus, for downstream analyses, an activity fraction cutoff of 0.7 was used to discard inactive library members unless stated otherwise. Most TdT variants in all 3 libraries pass this activity filter (99.8%, 91.6%, and 72.5% of deeply sequenced variants are active in libraries 1, 2, and 3, respectively). With both filters for activity fraction > 0.7 and insertions > 250 applied, 7193, 6640, and 6330 variants are represented in libraries 1, 2, and 3.

Plotting the average composition (A/T content) of insertions for every active TdT variant with at least 250 insertions revealed a wide spectrum of nucleotide preferences and average insertion lengths (Figure 3.1c). Apparent A/T biases (assuming A/T content in insertions ≅ A/T bias of the variant) range from 23% to 87%, and the average lengths of insertions vary from 1.4-4.9 bp (Figure 3.1c), resulting in a diverse collection of TdTs with various combinations of A/T biases and average insertion lengths. For our purposes, we wanted TdT variants with altered

biases, but essentially unchanged average insertion lengths. For other applications of TdT, however, it may be desirable to use variants with other specific combinations of these two properties; thus, we provide all performance metrics for every TdT variant in our libraries in Supplementary Tables 3-5. This constitutes a profile of 20,163 TdT variants (based on amino acid sequences) with activity fractions ≥ 0.7 and insertion counts ≥ 250. In addition to A/T content in insertion sequences, the complete database of these variants shows the breakdown of all 4 natural nucleotides inserted at the Cas9 target site. These data accurately reflect the sequences at the Cas9 target sites, but it should be noted that they don't necessarily reflect TdT's preferred substrates. This is because it was possible for TdT to insert nucleotides at the 3' end of the PAM-containing (top) strand, or at the 3' end of the target (bottom) strand after Cas9 created a DSB (Supplementary Figure 3.3). Whichever strand did not acquire an insertion directly from TdT would be provided with the complementary nucleotides during DNA repair, such that an insertion by TdT on the "bottom" strand would look like an insertion of the complementary base on the "top" strand. Therefore, we preferred to classify variants on the basis of A/T bias.

Because the molecular basis of TdT's natural bias for G's and C's is not well understood[2], we were interested in teasing out trends among the TdT variants with altered biases. Glancing at the scatterplots of A/T bias vs. average insertion length for each library shows the level of phenotypic diversity among variants in the libraries can be ranked as library 3 > library 1 > library 2 (Figure 3.1c). This suggests the target residues in library 3, on the alpha helix and pointing toward the nucleobase of the incoming ddNTP (PDB ID 4I27), are largely responsible for TdT's nucleotide preferences, but Loop1 also contributes to the biases. Another striking feature of the data is that many variants have exaggerated A/T-biases compared to wildtype TdT, whereas increased G/C-biased variants were rare. Wildtype TdT's preference for G's and C's has been previously noted as a departure from the behaviors of other polymerases that are occasionally forced to perform template-independent synthesis (*e.g.*, when polymerizing

across an abasic site), as they insert A's with ~100-fold higher efficiency than other nucleotides[2]. It is apparently easy to disrupt TdT's G/C bias and revert its nucleotide preferences to more closely match those of its A/T-preferring relatives. To examine the identities of amino acids responsible for altering the biases, the variants of each library were divided into 3 different bias bins (Figure 3.1c-d). When possible, the ~100 most G/C-biased mutants were placed in one bin for each library (library 2 had only 26 variants with an A/T bias < 0.5, so this bin is smaller than the others). Likewise, the ~100 variants with A/T biases closest to 0.5 (unbiased), and the ~100 most extremely A/T-biased variants in each library were sorted into separate bins. For the ~100 variants in each of these distinct bias bins, the amino acid sequences at the NNK loci were extracted. The composition of amino acids at each randomized position was calculated (*i.e.*, among the variants in the bin, what percentage have a given amino acid sequence at each NNK position), to determine if certain residues are over- or under-represented in each bias bin. To account for potential biases due to incomplete library coverage, the composition of amino acids in each bias bin was normalized to the composition of the full libraries, and to the composition of active TdT variants with at least 250 insertions (Supplementary Figure 3.4). The resulting heatmaps of amino acid compositions (Figure 3.1d) show D399, which was randomized in both libraries 1 and 2 (base of Loop1, and mid Loop1 libraries), is very strongly over-represented among G/C-biased variants, and non-wildtype residues are infrequent at this position in this bin. Also, D399 is completely absent from the A/T-biased bin in library 1 and is approximately neutral (not enriched or de-enriched) in the A/T-biased bin in library 2. This implies the wildtype D399 is an active player in bias determination. In library 3 (alpha helix near active site library), all 3 mutagenized positions (R454, E457, R461) frequently maintained their wildtype identities in the G/C-biased bin, and all 3 wildtype identities were depleted beyond detection in the A/T-biased bin. This suggests all 3 target codons in library 3 have an effect on nucleotide selection.

3.3.2 Properties of TdT variants determined in the multiplex assay are reproduced when assayed individually

To validate the results of the multiplex assay, TdT variants of interest were individually constructed and cloned into a mammalian expression vector. Overall, 75 library members were selected for individual characterization. They were chosen on the basis of their A/T biases and average insertion lengths, aiming for a diverse range in both criteria, while also prioritizing for high activity and number of insertions sequenced. Each TdT variant of interest was then co-transfected into HEK293T cells alongside a construct encoding Cas9 and a sgRNA targeting the genomic HEK3 site. Cell populations were collected 3 days later and editing outcomes at the HEK3 site were assessed *via* NGS. We observed a strong correlation between the variants' performance in the multiplex assay and the individual assays here for A/T bias (Pearson correlation coefficient, $\rho = 0.945$) and insertion length ($\rho = 0.917$) (Figures 3.2a and 3.2b). For A/T bias, the regression coefficient was 0.868 (95% confidence interval = 0.799-0.938) while for insertion length, the regression coefficient was 0.574 (95% confidence interval = 0.516-0.632), reflecting the observation that insertion lengths were shorter in the individual assays than in the multiplex assay (Figure 3.2b). This is likely due to the different methods for expressing TdT used in the multiplex assay (lentivirus) versus the individual assays (transient transfection). Because TdT is thought to use a distributive mechanism for polymerization[2], if lentiviral expression in the multiplex assay achieved higher intracellular concentrations of TdT, the over-expression could promote longer insertions. Nonetheless, the A/T bias and insertion length measurements from the multiplex assay can predict those of the individual assays, validating the multiplex assay as a way to isolate valuable TdT variants for application.

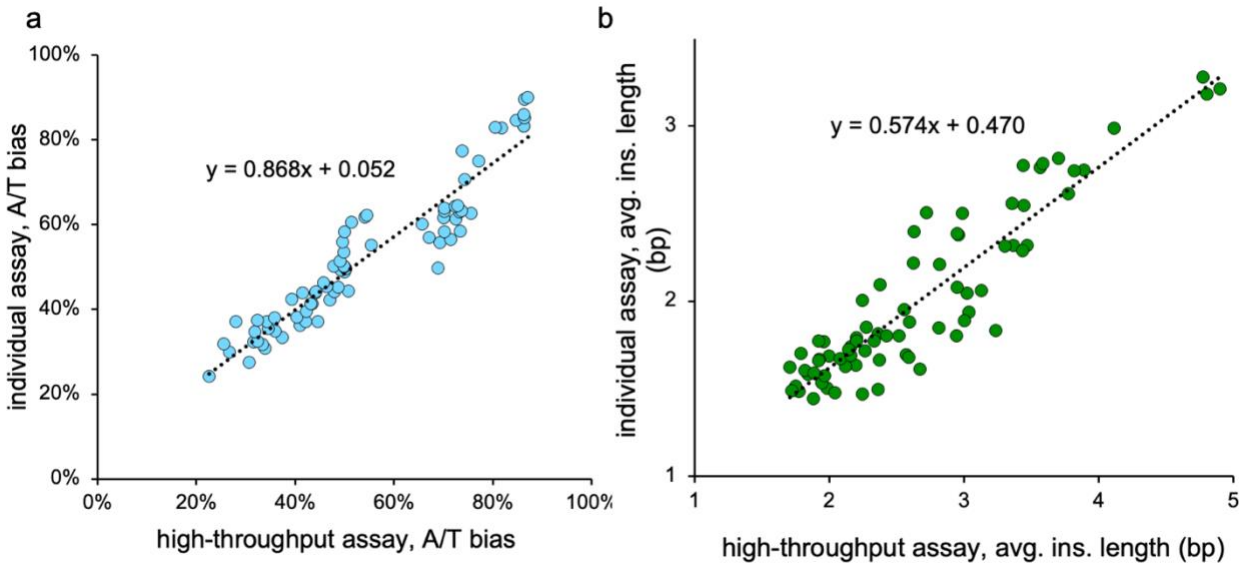


Figure 3.2. Comparison of TdT variant properties between the high-throughput assay and individual assays. **a,** 75 TdT variants were individually assayed after the high-throughput screen, and the apparent A/T biases of these variants (% of inserted nucleotides comprised of A/T) are plotted for both assay methods. **b,** The average insertion length for each of the 75 individually-assayed variants is shown for both assay methods.

3.3.3 *In vitro* biases of TdT mutants match those determined in the *in vivo* multiplex assay

When expressed in *E. coli* and assayed biochemically, the TdT variants recapitulated their A/T biases determined in the multiplex screen. We chose 7 TdT mutants to test *in vitro*, plus wildtype TdT for comparison. A TdT variant with nucleotide preferences nearly identical to wildtype was included to investigate if variants with similar performance *in vivo* also appear similar *in vitro*. All other TdT variants were selected because they had extreme biases. Two of the most A/T-biased variants from the pooled screen – A/T-biased1, with R454L, E457D, R461L and 86% A/T bias; and A/T-biased2, with R454I, E457K, R461A and 86% A/T bias – were included. Since dATP is the most disfavored nucleotide for wildtype TdT[2], we reasoned that these A/T-biased variants may be broadly useful for boosting the efficiency of *in vitro* DNA synthesis by TdT. Also, 3 unbiased variants – unbiased1, with R454M, E457C and 50% A/T bias; unbiased 2, with D396R, L398F and 50% A/T bias; and unbiased3, with R454K, E457R

and 50% A/T bias – were selected for the *in vitro* assay because equal incorporation of all bases would increase the entropy of TdT-generated sequences. We envisioned these could aid in synthesis of aptamer libraries[5]. In addition, G/C-biased variants were desirable subjects for *in vitro* testing, since G-biased TdTs may improve the performance of biosensors that synthesize G-quadruplexes for signal generation or enable the use of these biosensors in crude biosamples (without needing to purify specimens to remove all 4 natural nucleotides). However, because G/C-biased variants were rare among the library members we profiled in our multiplex assay, we attempted to amplify G/C bias by creating a double mutant possessing mutations from two modestly G/C-biased variants isolated from libraries 2 and 3 (L398M from library 2, plus R454D and E457D from library 3). We ultimately obtained one double mutant – G/C-biased1, with L398M, R454D, E457D – that was modestly more G/C-biased than its parent mutants, with an A/T fraction of 28% when tested individually in HEK293Ts (Supplementary Figure 3.5).

For all TdT variants, the overall biases were faithfully reproduced when assayed *in vitro*. There was a strong correlation between the A/T biases observed *in vitro* and *in vivo* (Pearson's $\rho = 0.950$), and the regression coefficient was 0.841 (95% confidence interval 0.564-1.118) (Figure 3.3a). We note that, because the *in vitro* assay profiled TdT's usage of all 4 nucleotides on a ssDNA primer instead of at a DSB, we gained additional information about how TdT variants prefer A vs. T and G vs. C. The *in vivo* multiplex screen cannot discern between direct TdT-mediated insertion of a nucleotide and copying of the inserted nucleotide due to DNA repair (Supplementary Figure 3). Therefore, the apparent rank order of each nucleotide's insertion frequency may be misleading in the *in vivo* assay results (as it was for a subset of the variants assayed here, Figure 3.3b), and we recommend *in vitro* assaying for applications where the ranking must be precisely known.

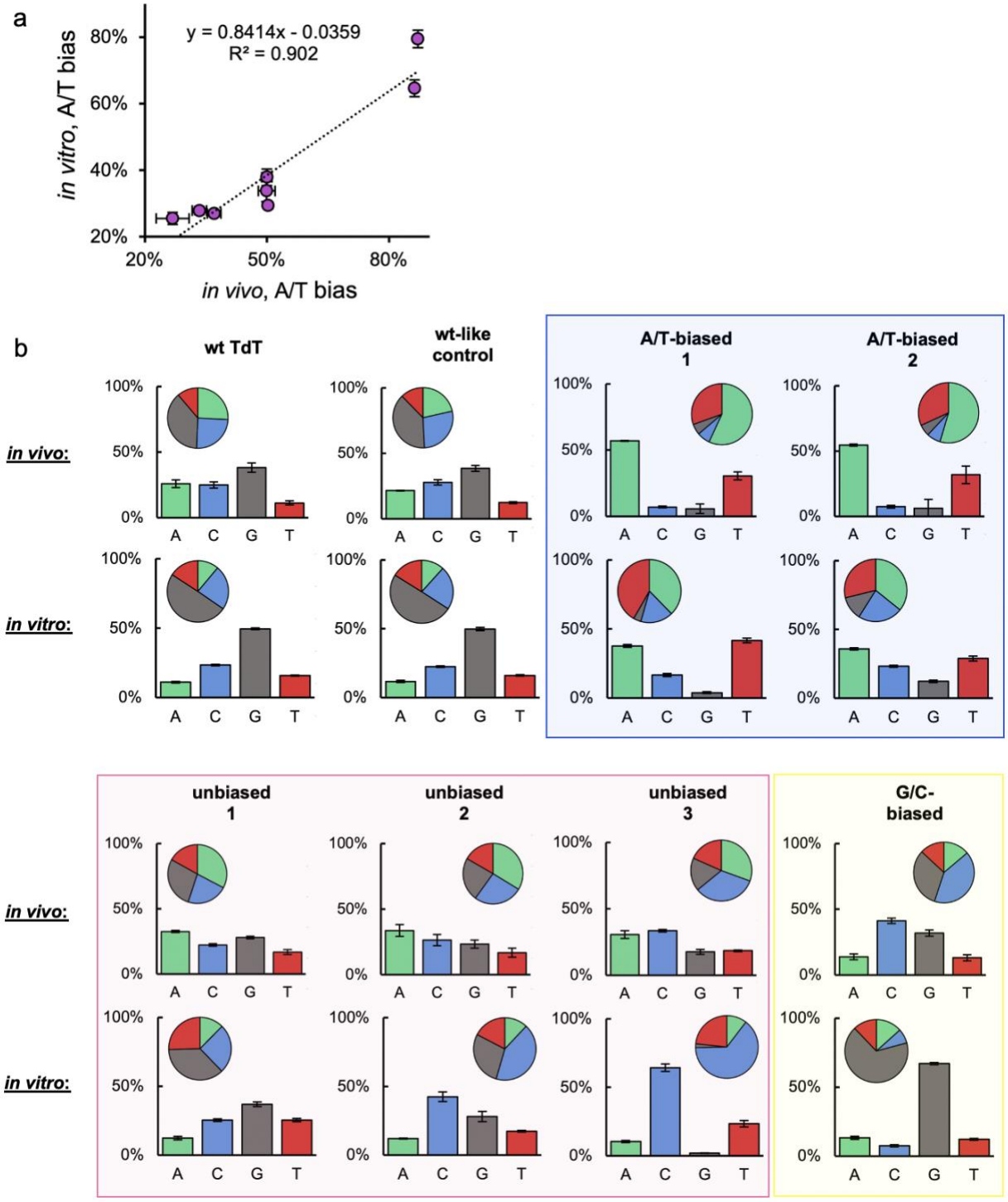


Figure 3.3. TdT mutants with practically useful nucleotide preferences (*i.e.*, having strong biases or equal affinity for A/T and G/C) were assayed *in vitro*. **a,** Comparison between the apparent A/T biases of TdT variants *in vivo* and in the individual assays *in vitro*. Except for the G/C-biased variant, *in vivo* data is from the multiplex assay. The G/C-biased variant was assayed individually *in vivo*. **b,** Bar and pie graphs both depict the % of incorporated nucleotides comprised of A, C, G, or T for each variant, when tested *in vivo* and *in vitro*. Error bars indicate the standard deviation for 3 technical replicates of the high-throughput assay, and 3 biological replicates for the *in vitro* assay. 3 biological replicates are shown for the *in vivo* assay of the G/C-biased variant.

3.3.4 Engineered TdTs improve barcode diversity for DNA recording in live cells

Cell HistorY Recording by Ordered iNsertion (CHYRON) is a molecular memory recorder that our group has previously established to enable the study of tissue development in a longitudinal, non-disruptive fashion (Figure 3.4a)[22]. Briefly, the mechanism of CHYRON recording is the continuous accumulation of random insertion mutations at a neutral locus in the genome. These insertions can be used to record the lineage relationships among cells in live tissue, since the random sequences (barcodes) can uniquely label unrelated cells. Descendants of the barcoded cells inherit an exact replica of the barcode from their progenitors, such that shared barcodes among cells effectively mark the related cells in a tissue. Furthermore, because the barcodes continuously aggregate random insertions, cells that are distantly related (and therefore inherited a barcode from a shared progenitor early on) are likely to gain additional insertions in their barcodes, thereby marking their separate clades in the lineage tree. Iterative accumulation of insertions is possible because the locus that stores CHYRON records is the gene for a homing guide RNA (hgRNA)[29,35]. CRISPR-Cas9 and TdT are co-expressed to cut the hgRNA locus and acquire an insertion (~3 bp on average) at the resulting double strand break (DSB). The mutated hgRNA is transcribed, so it can direct Cas9 back to the same locus for the next round of editing. The original version of CHYRON used wildtype TdT for insertion writing, but we and others[2] observed that the wildtype polymerase has a definitive bias to utilize dGTP or dCTP. Because of this bias, the full spectrum of possible barcode sequences isn't readily generated at the CHYRON loci. This increases the probability of unrelated cells receiving identical barcode sequences, resulting in the false appearance of shared ancestry (homoplasy). To minimize this risk of homoplasy, we turned to the unbiased TdTs discovered in our high throughput screen. We hypothesized that we could directly substitute them into the CHYRON framework and generate truly random insertion sequences with unprecedently high information content.

Unbiased TdTs were selected from our database of variants from the multiplex screen, then installed into CHYRON cell lines *via* genomic integration. From the collection of variants

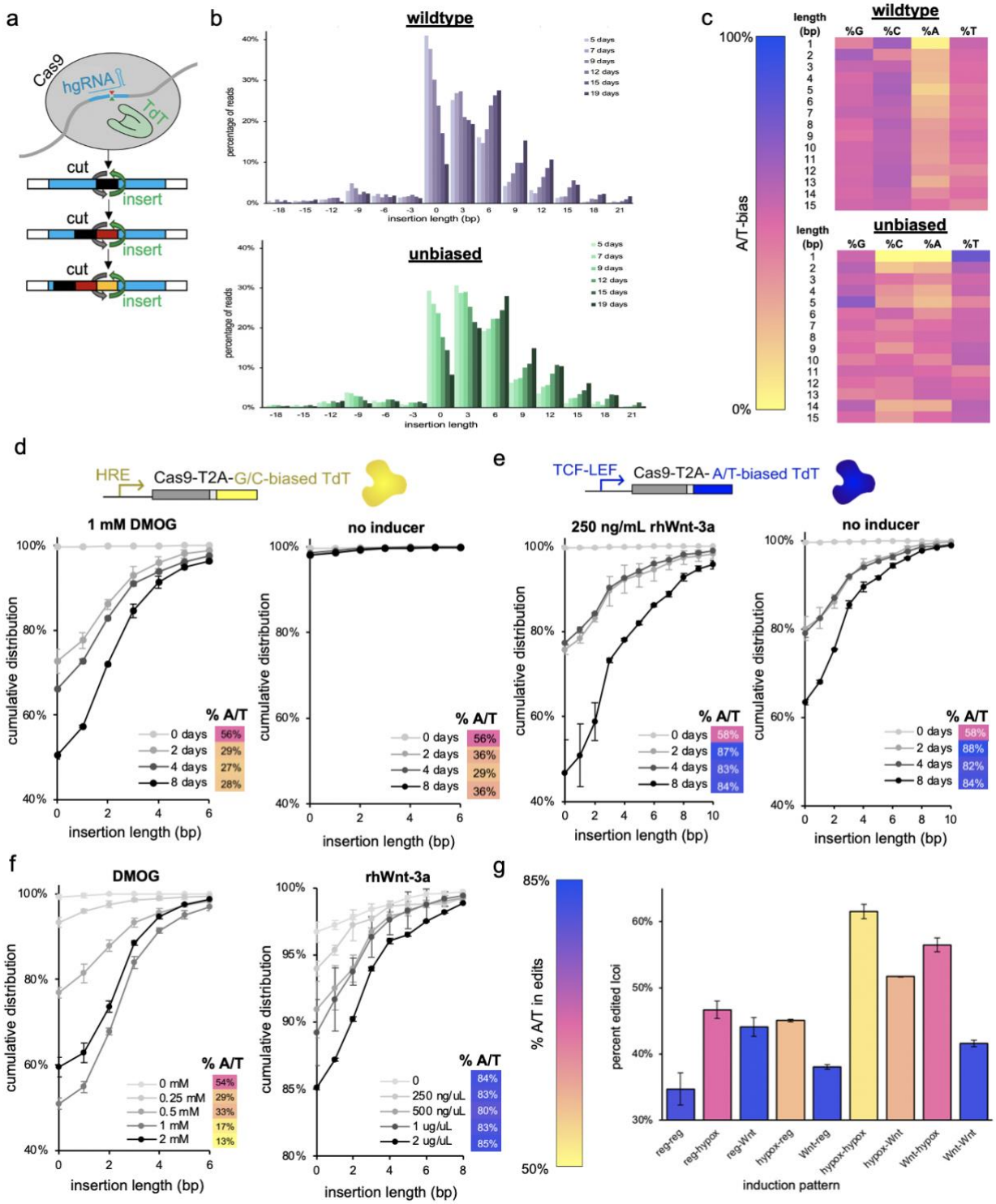


Figure 3.4. Altered-bias TdTs were implemented for DNA recording in live HEK293T cells. **a,** Depiction of CHYRON, an *in vivo* DNA recorder that iteratively acquires insertion mutations through the joint activity of Cas9, a homing guide RNA (hgRNA), and TdT. **b,** CHYRON was propagated for 19 days with either wildtype TdT or an unbiased TdT. The length of insertions were analyzed at various timepoints to compare editing efficiency between the two TdT variants. **c,** The composition of insertions at the CHYRON locus at the end of the 19-day experiment is shown for wildtype and unbiased TdTs. **d,** CHYRON with a G/C-biased TdT that is driven by a hypoxia-inducible promoter was induced for various durations with a chemical hypoxia mimic (DMOG). **e,** CHYRON with an A/T-biased TdT that is driven by a Wnt-inducible promoter was induced for various durations with a recombinant Wnt protein (rhWnt-3a). **f,** Different concentrations of DMOG and Wnt were used to induce recording with the corresponding inducible CHYRON cassettes. **g,** Both hypoxia- and Wnt– inducible CHYRON cassettes with differently-biased TdTs were simultaneously introduced into populations of HEK293Ts, which were then exposed to various patterns of induction. Error bars depict the standard deviations among 2 technical replicates of each data point.

with A/T biases in the range of 45-55% A/T, we narrowed down our selection by checking for approximately equal proportions of A, C, G, and T inserted across all insertion lengths. In addition to the composition of nucleotides, the frequency of successful insertions at a Cas9 cut site ("activity fraction") was considered, since variants with low activity would diminish the overall performance of CHYRON. Ultimately, 3 unbiased mutants with activity fractions ≥ 0.8 (ins / ins+del) were selected. These are the same variants that were assayed *in vitro*, Unbiased1, Unbiased2, and Unbiased3 (Figure 3.3). Stable cell lines were established using Cas9 and homology directed repair to integrate a single hgRNA gene per genome at AAVS1, followed by piggyBAC transposition to integrate pCAG-Cas9-T2A-TdT genes. These cell lines were continuously passaged for 19 days. Periodically, subpopulations were sequenced, and their distributions of indel sizes and nucleotide compositions were quantified.

The most beneficial variant, which achieved near-perfect barcode diversity, was found to propagate the CHYRON locus with similar efficiency to wildtype TdT. Unbiased3 was the frontrunner in this propagation timecourse, but all of the unbiased candidates improved barcode diversity relative to wildtype TdT and propagated the recording locus efficiently. The efficiency of editing was apparent in the continuous shift toward edited loci (reads with an insertion length of 0 diminished over time), and the gradual accumulation of long insertions (indicative of multiple rounds of editing) that occurred at nearly identical rates for wildtype and unbiased TdTs (Figure 3.4b). Barcode diversity was calculated in terms of the Shannon Entropy (a measure of

unpredictability) at the end of the 19-day experiment and compared to wildtype TdT. Wildtype TdT's insertions reflected an overall A/T bias of 63%, and a corresponding Shannon Entropy of 1.93 bits per inserted nucleotide. On the other hand, the unbiased TdT achieved 1.99 bits per inserted nucleotide (compared to 2 bits for the theoretical maximum). Thus, instances of homoplasy due to under-utilization of the available barcode building blocks is practically eliminated with this near-perfect unbiased TdT. The next step to eliminate homoplasy altogether will be increasing the efficiency of CHYRON recording, so that longer barcodes are generated on average. However, this is beyond the scope of the current study.

3.3.5 TdT mutants with extreme nucleotide insertion preferences enable dual-channel signal recording in HEK293T cells

In our previous work, we used our *in vivo* DNA recorder, CHYRON (Figure 3.4a), as an analog recorder for hypoxia[22]. We demonstrated that the combined activities of a hypoxia-inducible Cas9, hypoxia-inducible TdT, and constitutively expressed homing guide RNA (hgRNA) produce a hypoxia-dependent increase in the number of insertion mutations in the recording loci. Different doses and durations of exposure to a chemical hypoxia mimic (DMOG) led to differing degrees of mutagenesis at the recording loci in the populations of CHYRON-expressing cells. The hypoxia-recording experiments served as a first proof of principle for analog signal recording, but in theory, CHYRON could be used to record any event that causes a transcriptional response in the cells of interest. Provided that the corresponding transcription factor binding sites are known, recording the desired signal only requires substituting the appropriate inducible promoter into the Cas9 and TdT expression cassette. Thus, CHYRON has the potential to record many different signals, but critically, it can only record one signal at a time. With wildtype TdT, the identities of the nucleotides in the insertion mutations are uninformative, while the extent of editing provides the desired information (how much inducer / how long the cells were exposed to inducer). We aimed to alter this paradigm by imbuing the

identities of the insertion mutations with meaning. Specifically, we used an A/T-biased TdT variant to record one signal of interest (Wnt), and a G/C-biased TdT variant to record a second signal of interest (hypoxia). With this setup, we envisioned that both the extent and identities of insertions could be used to decipher the cells' histories of induction.

Strongly biased TdT variants were selected as follows: candidate TdTs needed to 1) have extreme preference for adding either A/T or G/C, 2) have high activity fractions (when similarly biased variants were considered, the option with higher activity were selected), and 3) make relatively short insertions in each round of editing (~3 bp), since CHYRON's editing efficiency declines when the hgRNA grows too long due to insertion mutagenesis. For the A/T-biased TdTs, we additionally considered the observed frequency of adding not only T's, but also A's, since the recording locus needs to be continuously transcribed to produce functional hgRNAs, and homopolymers of 4 or more Ts would constitute a premature transcription termination signal in the hgRNA gene (4xT's terminates transcription with approximately 75% efficiency, and 100% efficient termination is achieved with 6xT's[36]). We selected 1 A/T-biased TdT, and 1 G/C-biased TdT. These are the same variants that were assayed *in vitro* (Figure 3.3), A/T-biased2 (86% A/T), and G/C-biased (28% A/T). The variants were cloned into inducible cassettes with the general architecture [inducible promoter]-Cas9-T2A-TdT, so that Cas9 would also be expressed only in response to the inducer of interest.

Before testing the biased TdTs' ability to perform dual-channel recording, we first assessed their individual performances as analog recorders, in separate populations of cells. The recording cassettes were transiently transfected into HEK293Ts that already possessed a constitutive hgRNA gene stably integrated in the genome (to enable iterative rounds of editing). These populations of cells were either exposed to the relevant inducer or passaged in inducer-free media and sequenced with Illumina at various timepoints over the course of 8 days. Encouragingly, when the hypoxia-inducible cassette with a G/C-biased TdT was induced with DMOG, the cumulative distribution of insertion lengths was clearly distinguishable for different

durations of induction (Figure 3.4d). This shows the overall percent of edited CHYRON loci and the abundance of long insertions both increased proportionally to duration of DMOG exposure. The A/T content of the insertions at the CHYRON loci are constant over time, at 27-29% A/T after 2+ days of induction. This is consistent with the known bias of the G/C-biased TdT. When the same cell line was passaged in regular media (no inducer), there was almost no leaky editing from 0-8 days. A/T content in the few insertion mutations due to leaky expression is slightly higher than in the induced samples (29-36% A/T), which attests to the level of TdT expression being very low when no inducer is present; therefore, Cas9-generated DSBs are less frequently edited by TdT. When TdT misses an opportunity to edit a DSB, NHEJ often results in small insertions with ~50% A/T content, skewing the bias away from that of the specialized TdT. Overall, the behavior of the hypoxia-recording cassette allowed different durations of DMOG exposure to be classified from the CHYRON records, and the presence or absence of inducer was clear based on the extent of editing. When the same experiment was conducted for the Wnt-inducible cassette with an A/T-biased TdT (Figure 3.4e), the resolution of recording was somewhat coarse, as 2-4 days of induction have indiscernible cumulative distributions of insertion lengths. However, 0, 2-4, and 8 days all produced distinct compositions of edit lengths, so we consider the resolution of this recorder to be approximately 4 days. The A/T content of the insertions was consistent with the expected values for this A/T-biased variant (84-87% at 2+ days of induction). When an identical population of cells was passaged in the absence of inducer, a considerable amount of leaky editing occurred. This complicates the interpretation of recorded insertions, but it doesn't preclude the use of this Wnt recorder. Editing proceeded at a slower pace in the absence of inducer (for each timepoint, the induced population was more extensively edited than the uninduced population), so the relative amounts of editing between populations can be used to draw conclusions about which population experienced a longer period of induction. Compositions of insertions in the uninduced population are approximately the same as the induced population, since the A/T-biased TdT

appears to be expressed at a high level in the absence of inducer (82-88% A/T after 2+ days of passaging). Next, we tested the two distinct recorders, again in separate populations, but this time for their responsiveness to different doses of inducer (Figure 3.4f). Populations of HEK293Ts were transfected with an inducible cassette, induced with a range of inducer concentrations, then collected and sequenced *via* NGS after 2.5 days of induction. The dose responses of both recorders exhibited almost ideal trends, with increased editing (both overall percent edited loci and length of insertions) in the presence of higher inducer concentrations. There was one exception to the trend, however, as the hypoxia recorder produced less editing with 2 mM DMOG than with 1 mM DMOG in the culture media. We believe this was due to the harmful health effects of DMOG on the cells, as we observed high rates of cell death in the cultures with high DMOG concentrations. In this harsh environment, the combined cytotoxicity of DMOG and Cas9 expression could lead to editing-competent cells being selected against. The A/T content of the insertions in the hypoxia recording experiment trend toward decreased A/T usage in response to increased inducer (29% A/T for 0.25 mM DMOG, compared to 13% A/T for 2 mM DMOG). This implies the expression level of the G/C-biased TdT falls in a range where the effect of NHEJ-mediated insertions can be seen with low levels of induction. With high levels of induction, however, the G/C-biased TdT is expressed at a higher level and therefore outcompetes the NHEJ machinery more frequently, creating G/C-rich insertions. Therefore, the A/T content of the insertions reflects the dose of inducer, even when the hypoxia recorder is the only recorder in the population. In the case of the Wnt recorder, the A/T content of the insertions fall in line with the expected values for high expression of the recorder (80-85% A/T).

Having confirmed the hypoxia recording cassette can generate G/C-rich insertions, and the Wnt recording cassette can generate A/T-rich insertions in an analog fashion, we next performed dual-channel recording of both hypoxia and Wnt in a single population of HEK293Ts. For this, populations of cells were co-transfected with both recording cassettes, then induced

with hypoxia, Wnt, or nothing for 4 days (1 “epoch”). Then, the populations were re-transfected with the recording cassettes and subjected to an additional epoch of induction with hypoxia, Wnt or nothing. The final populations were sequenced on an Illumina platform, and compositions of insertions at the CHYRON loci were inspected. Knowing that the Wnt recorder is leaky (Figure 4e-f), the presence of A/T-rich insertions (82%) in the population that experienced 0 epochs of induction is unsurprising. Yet, the lack of induction could still be identified by the low extent of editing. Generally, the percent of CHYRON loci bearing an edit correlated with the amount of time the populations were induced. For 8/9 populations, the total percent of CHYRON loci with an edit followed the trend 0 epochs < 1 epoch < 2 epochs of induction. In addition, 9/9 populations accurately recorded the identities of the inducers in each epoch, as reflected in the A/T content of the CHYRON records. Every sample that exclusively experienced Wnt induction had insertions comprised of ~83% A/T. On the other end of the spectrum, two epochs of hypoxia induction produced more G/C-rich insertions (53% A/T), as expected. Although the G/C-rich TdT used to record hypoxia has an A/T bias of 28%, the final composition is 53% A/T because the leaky Wnt recorder competed to make insertions and was moderately successful in doing so. Samples exposed to both hypoxia and Wnt, or hypoxia and regular media, generated CHYRON records with moderate A/T content (59-73% A/T). The known leakiness of the Wnt recorder, and lack of leakiness in the hypoxia recorder, explains the similar A/T content in samples that were only induced with hypoxia for 1 epoch. When hypoxia was induced for 1 epoch, the hypoxia recorder likely generated insertions with ~53% A/T content for that single epoch (if they behaved the same as the sample that was induced with hypoxia for the entire experiment). Upon Wnt induction, or no induction in the other epoch, the hypoxia recorder was effectively turned off, while the Wnt recorder added A/T-rich insertions to the recording loci, regardless of if it was induced or just leakily expressed (bringing up the final % A/T). While inspecting the A/T biases of the records, we also noted that the populations exposed to hypoxia for only 1 epoch had a lower A/T content if hypoxia was the first epoch (59-61% A/T, compared

to 71-73% A/T if hypoxia was the second epoch). This is likely because the hgRNA in this instantiation of CHYRON was the optimal length (20 nt) for efficient Cas9 cutting at the beginning of the experiment. Therefore, in the first epoch, the entire population of cells possessed a hgRNA that was conducive to high editing rates, and many insertions were created with an A/T content indicative of the inducer. In the second epoch, the population contained a mixture of cells that had already acquired an insertion and therefore had sub-optimal hgRNAs, and a sub-population of unedited cells that still had an optimal hgRNA. This produced the conditions for a relatively high rate of editing in the first epoch (so the first inducer led to many edits), and a lower rate of editing in the second epoch (so the second inducer yielded fewer edits and had a lesser impact on the overall A/T content in the records). In this way, the order of inducers is encoded in the CHYRON records. Although the order of inducers could theoretically be deduced with a mathematical model that accounts for the change in efficiency over rounds of editing, it would be preferable to engineer CHYRON to maintain a constant rate of editing in the future (*e.g.*, co-opting a ribonuclease to remove excess nucleotides from the spacer). Also, future efforts should focus on reducing leaky CHYRON recording (*e.g.*, fusing degrons to Cas9 and TdT). Since we want to enable recording of any signal with a known inducible promoter, we need tools to refine the precision of editing despite leaky basal expression.

3.4 Discussion

Here, we present a new approach to screening libraries of TdT variants with unprecedently high throughput. Using this screening method, we successfully identified TdTs with advantageous properties for our application of interest (*in vivo* DNA recording with CHYRON[22]) and simultaneously generated a database describing the activities, insertion biases, and average insertion lengths of 23,180 TdT variants. Although the high-throughput screen was conducted in HEK293T cells, we demonstrated that the insertion biases *in vivo* are conserved *in vitro*. The TdT variants we have already characterized, and the massively parallel screen we developed,

could both aid the development of other *in vivo* applications and *in vitro* applications of TdT (of which there are many).

Among the TdT variants screened in this study, we captured a wide variety of nucleotide preferences (propensity to insert A/T vs. C/G) and average insertion lengths *in vivo* (suggesting altered K_m and/or k_{cat}). We imagine some of these variants could be directly installed into other applications for technological gains. A straightforward example of this would be to implement the same altered-bias variants we used to improve CHYRON (the Unbiased2, A/T-biased2, and G/C-biased variants) in an alternative *in vivo* DNA recorder called DARLIN [37]. Instead of iteratively editing a single recording locus, the DARLIN architecture uses Cas9 plus TdT to generate random insertions at many recording sites across the genome, and it has been proven to yield high barcode diversity while also allowing simultaneous lineage barcode, transcriptome, and epigenome readout from single cells. Substituting wildtype TdT with Unbiased2 could further increase the barcode diversity, thereby increasing the number of cells that can be individually barcoded, and broadening the range of developmental processes that could be studied with DARLIN. Alternatively, implementing dual-channel signal recording for DARLIN with A/T-biased2 and G/C-biased variants could allow population-level recording of cells' exposures to two signals of interest (did the population experience one signal which induces A/T-rich insertions, the other which induces G/C-rich insertions, or both? The final A/T content of insertions could report on this). As for *in vitro* applications, TdTs that generate longer insertions on average are promising candidates for stepwise oligonucleotide synthesis, or any application that seeks high activity. While detailed biochemical characterization is still needed, it is feasible that longer average insertions indicate faster catalytic turnover. The A/T-biased, fast inserting TdTs we found (A/T-biased1 and A/T-biased2) could potentially be substituted into enzymatic oligo synthesis platforms to overcome the currently slow rate of dATP addition (dATP requires the longest coupling time out of all 4 natural nucleotides, 2x longer than the others[18]). Other appealing applications for TdT variants with improved kinetics include biosensing (faster results)

and DNA tailing (shorter protocols). On the other end of the spectrum, variants that make shorter insertions on average, but are still active, may have utility in DNA data storage. Recent systems for DNA data storage allow TdT to insert homopolymeric runs of each base, and the transitions between base identities are used to encode information[16]. Apyrase is currently used to limit the length of homopolymers by degrading dNTPs during *in vitro* TdT extension reactions, but a cheaper option would be to use TdT variants that are engineered to either have lower affinity for the substrates or slower turnover. Similarly, libraries of aptamers synthesized by TdT require control over the approximate length of the aptamers [5], which could be provided by our short-inserter TdTs.

For applications that demand new properties beyond those of the TdT variants in this study, the general approach we developed for testing tens of thousands of library members in parallel could facilitate new engineering endeavors. When deciding which engineering goals are feasible, it is important to note the behavior of each variant was read out *via* NGS in our screen for altered-bias variants. This necessitated PCR amplification of the site harboring TdT-mediated insertions, and base pairing between the PCR amplicons and Illumina's sequencing-by-synthesis probes. Therefore, the most straightforward goals to pursue with this screen are those that involve incorporating natural nucleotides, or modified nucleotides that are compatible with PCR and NGS (*e.g.*, making TdTs that make even longer insertions on average in the quest for improved kinetics, or making TdTs that exhibit stronger shifts in nucleotide biases when exposed to different divalent cations, for DNA data storage as in TURTLES[17]). Alternatively, instead of sequencing the inserted nucleotides, enrichment could be used to identify successful TdT variants for some applications. For example, TdT is often used to label DNA with modified nucleotides that allow the DNA to be captured (*e.g.*, with biotin or His-tag mimics). Libraries designed to improve incorporation efficiency of these modified nucleotides could be tested *in vivo*, then the TdT-mediated insertions could be enriched with streptavidin, Ni^{2+}- and Eu^{3+}-NTA-agarose solid supports, or other enrichment methods. Because the TdT-

mediated insertions are on the same molecule as the corresponding TdT variant's gene, the TdT genes that are enriched would represent the effective variants, without needing to sequence the insertions. Also, aside from directly screening for incorporation of the desired substrate, the high-throughput screen could be used to test if TdTs are still functional after receiving mutations that are designed to remodel the protein in a beneficial way (*e.g.*, efforts to increase the size of the active cavity have improved TdT's activity with 3'-O-blocked nucleotides for stepwise oligonucleotide synthesis [20], and perhaps further active cavity expansion could yield even greater efficiencies). The *in vivo* screen could efficiently test tens of thousands of prospective designs and narrow down the library of TdT variants to a reasonable number for subsequent *in vitro* characterization with the substrate of interest.

3.5 References

1. Ashley, J., Potts, I. G. & Olorunniji, F. J. Applications of terminal deoxynucleotidyl transferase enzyme in biotechnology. *ChemBioChem* **24**, e202200510 (2023).
2. Motea, E. A. & Berdis, A. J. Terminal deoxynucleotidyl transferase: the story of a misguided DNA polymerase. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1804**, 1151–1166 (2010).
3. Li, L., Bowling, S., Wang, S. & Camargo, F. 2015–A MOUSE MODEL WITH HIGH CLONAL BARCODE DIVERSITY FOR JOINT LINEAGE, TRANSCRIPTOMIC, AND EPIGENOMIC PROFILING IN SINGLE CELLS. *Exp Hematol* **124**, S44 (2023).
4. Li, L. *et al.* A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells. *Cell* **186**, 5183–5199 (2023).
5. Ashley, J. *et al.* Terminal deoxynucleotidyl transferase-mediated formation of protein binding polynucleotides. *Nucleic Acids Res* **49**, 1065–1074 (2021).

6. Huang, Z. *et al.* Terminal deoxynucleotidyl transferase based signal amplification for enzyme-linked aptamer-sorbent assay of colorectal cancer exosomes. *Talanta* **218**, 121089 (2020).
7. Du, Y.-C., Zhu, Y.-J., Li, X.-Y. & Kong, D.-M. Amplified detection of genome-containing biological targets using terminal deoxynucleotidyl transferase-assisted rolling circle amplification. *Chemical Communications* **54**, 682–685 (2018).
8. Zhao, B., Gong, Z., Ma, Z., Wang, D. & Jin, Y. Simple and sensitive microRNA labeling by terminal deoxynucleotidyl transferase. *Acta Biochim Biophys Sin* **44**, 129–135 (2012).
9. Chen, X. *et al.* Terminal deoxynucleotidyl transferase induced activators to unlock the trans-cleavage of CRISPR/Cpf 1 (TdT-IU-CRISPR/Cpf 1): an ultrasensitive biosensor for Dam MTase activity detection. *Biosens Bioelectron* **163**, 112271 (2020).
10. Li, Z. *et al.* Label-free fluorescence strategy for methyltransferase activity assay based on poly-thymine copper nanoclusters engineered by terminal deoxynucleotidyl transferase. *Spectrochim Acta A Mol Biomol Spectrosc* **260**, 119924 (2021).
11. Abnous, K. *et al.* A highly sensitive electrochemical aptasensor for cocaine detection based on CRISPR-Cas12a and terminal deoxynucleotidyl transferase as signal amplifiers. *Talanta* **241**, 123276 (2022).
12. Zhang, X. *et al.* CRISPR-Cas12a coupled with terminal deoxynucleotidyl transferase mediated isothermal amplification for sensitive detection of polynucleotide kinase activity. *Sens Actuators B Chem* **330**, 129317 (2021).
13. Li, X. *et al.* Ultrasensitive detection of microRNAs based on click chemistry-terminal deoxynucleotidyl transferase combined with CRISPR/Cas12a. *Biochimie* **208**, 38–45 (2023).
14. Schneider, L. A., Sauter, B., Dagher, K. & Gillingham, D. Recording Binding Information Directly into DNA-Encoded Libraries Using Terminal Deoxynucleotidyl Transferase. *J Am Chem Soc* **145**, 20874–20882 (2023).

15. Land, H., Grez, M., Hauser, H., Lindenmaier, W. & Schütz, G. 5′-Terminal sequences of eucaryotic mRNA can be cloned with high efficiency. *Nucleic Acids Res* **9**, 2251–2266 (1981).

16. 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* **10**, 2383 (2019).

17. Bhan, N. *et al.* Recording temporal signals with minutes resolution using enzymatic DNA synthesis. *J Am Chem Soc* **143**, 16630–16640 (2021).

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

19. Chua, J. P. S. *et al.* Evolving a thermostable terminal deoxynucleotidyl transferase. *ACS Synth Biol* **9**, 1725–1735 (2020).

20. Lu, X. *et al.* Enzymatic DNA synthesis by engineering terminal deoxynucleotidyl transferase. *ACS Catal* **12**, 2988–2997 (2022).

21. Barthel, S., Palluk, S., Hillson, N. J., Keasling, J. D. & Arlow, D. H. Enhancing terminal deoxynucleotidyl transferase activity on substrates with 3′ terminal structures for enzymatic de novo DNA synthesis. *Genes (Basel)* **11**, 102 (2020).

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

23. Tsai, S. Q. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat Biotechnol* **32**, 569–576 (2014).

24. Fonseca, J. P. *et al.* A toolkit for rapid modular construction of biological circuits in mammalian cells. *ACS Synth Biol* **8**, 2593–2606 (2019).

25. Ede, C., Chen, X., Lin, M.-Y. & Chen, Y. Y. Quantitative analyses of core promoters enable precise engineering of regulated gene expression in mammalian cells. *ACS Synth Biol* **5**, 395–404 (2016).

26. Veeman, M. T., Slusarski, D. C., Kaykas, A., Louie, S. H. & Moon, R. T. Zebrafish prickle, a modulator of noncanonical Wnt/Fz signaling, regulates gastrulation movements. *Current biology* **13**, 680–685 (2003).

27. Mahajan, K. N. *et al.* Association of terminal deoxynucleotidyl transferase with Ku. *Proceedings of the National Academy of Sciences* **96**, 13926–13931 (1999).

28. Cha, R. S. & Thilly, W. G. Specificity, efficiency, and fidelity of PCR. *Genome Res* **3**, S18–S29 (1993).

29. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science (1979)* **353**, aag0511 (2016).

30. Nov, Y. When second best is good enough: another probabilistic look at saturation mutagenesis. *Appl Environ Microbiol* **78**, 258–262 (2012).

31. Clopper, C. J. & Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404–413 (1934).

32. Schaudy, E., Lietard, J. & Somoza, M. M. Sequence preference and initiator promiscuity for De Novo DNA synthesis by terminal deoxynucleotidyl transferase. *ACS Synth Biol* **10**, 1750–1760 (2021).

33. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science (1979)* **343**, 84–87 (2014).

34. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science (1979)* **343**, 84–87 (2014).

35. Kalhor, R., Mali, P. & Church, G. M. Rapidly evolving homing CRISPR barcodes. *Nat Methods* **14**, 195–200 (2017).

36. Gao, Z., Herrera-Carrillo, E. & Berkhout, B. Delineation of the exact transcription termination signal for type 3 polymerase III. *Molecular Therapy-Nucleic Acids* **10**, 36–44 (2018).

37. Li, L. *et al.* A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells. *Cell* **186**, 5183–5199 (2023).

CHAPTER 4: prime editing CHYRON (peCHYRON) is a next-generation DNA recorder that overcomes many of the limitations of CHYRON

4.1 Introduction

Understanding and eventually engineering development will require scientists to relate both cell lineage relationships and patterns of signaling to specific cell fates. To achieve this goal, it will first be necessary to deconvolute the phylogenies of individual cells, and also follow many signaling pathways in parallel, over long timescales, in single cells, as those cells differentiate. DNA recording can make these analyses possible by enabling each cell to record transient phenomena (*e.g.*, the existence of an ancestral stem cell that undergoes apoptosis after giving rise to differentiated daughter cells, or the presence of fleeting microenvironmental cues that affect gene expression within the cell of interest) and permanently store a record of the transient event as a mutation in its DNA (Figure 4.1a). To maximize the amount of information that can be extracted from its records, a DNA recorder must have the following properties: (1) high information content: the ability to generate records with diverse sequences, either random sequences that can be used to uniquely barcode individual cells and therefore mark distinct cell lineages with high confidence, or sequences that encode patterns of transient events; (2) order: the ability to record events in the sequence they arrive, so the timing of events can be deciphered from the relative positions of mutations in the records; (3) punctuation: the ability to delimit one event from the next in the record; (4) durability: the ability to record new events without corrupting or deleting the record of older events; (5) continuous operation: the ability to record new events over long enough timescales to study biologically interesting processes; and (6) multiplexability: the ability to record multiple distinct signals in parallel.

Before the invention of prime editor[1], DNA recorders suffered from fundamental mechanistic tradeoffs that prevented them from achieving all these properties simultaneously[2–14,16–19]. These recorders either caused double-strand breaks in DNA, compromising durability because corruptive mutations would arise during DNA repair, or they could install only a limited number of potential mutations (*e.g.*, base editors). However, DNA recorders that use prime editor, a nickase Cas9-reverse transcriptase (RT) fusion that installs precise mutations specified by pegRNAs[15,22], do not suffer from these same limitations. As a result of the characteristics of prime editor, and the fundamental design of peCHYRON, our recorder creates an ordered array of high-information edits that can be extended for an arbitrarily high number of rounds in principle (Figure 4.1b).

4.2 Methods

4.2.1 Plasmid cloning

Polymerase chain reactions (PCRs) were performed with Q5 Hot Start High-Fidelity DNA Polymerase (NEB M0493) or Phusion Hot Start Flex DNA Polymerase (NEB M0535). PCR reagents and all DNA-modifying enzymes were also purchased from NEB, and all primers were synthesized by Integrated DNA Technologies (IDT). Plasmids were made by standard Gibson assembly[31], *in vivo* recombination, Golden Gate assembly, or NEBuilder HiFi assembly (NEB E2621). Specifically, pegRNA-expressing plasmids were made either by adding or changing sequences on PCR primers, then performing *in vivo* recombination in Top10 bacteria, or by ordering pegRNA sequences as long oligos from IDT, then performing NEBuilder HiFi assembly. PiggyBac cargo plasmids were assembled by Golden Gate. All plasmids to be used for transfection were purified from Top10 or NEBStable *E. coli* (NEB C3040) with HP GenElute Midi or Mini kits (Sigma # NA0200 and NA0150).

Plasmids encoding PE2 (Addgene plasmid #132775[21]), PEmax (Addgene plasmid #174820[23]), and AncBE4Max (Addgene plasmid #112094[32]) were a gift from David Liu (Broad Institute). PiggyBac cargo plasmids used in Figures 4.3 and 4.5 and pegRNA-expressing plasmids used in Figure 4.4 were created using the Mammalian Toolkit (Addgene article #28197510[33]), which was a gift from Hana El-Samad (UCSF). The sequences of TetR and LacI (Addgene plasmid #81253[16]) were a gift from Timothy Lu. Doxycyline- and IPTG-inducible U6 promoters were initially adopted from the same publication[16]. The IPTG inducible U6 promoter included in that publication consisted of one LacO1 and one LacOS sequence in the promoter and an additional LacOS sequence downstream of the guide RNA. We initially used LacO1 and LacOS in the promoter, but eventually used a promoter with two internal LacOS sequences. For chemical inducer recording, we also used constitutive U6 promoters with mutations we added to reduce their activity. Annotated full plasmid maps are available at github.com/liusynevolab/peCHYRON. Pools of PiggyBac cargo plasmids that can be used to make peCHYRON cell lines will be made available from Addgene.

4.2.2 Cell culture and transfection

HEK293T (CRL-3216), U-2 OS (HTB-96), K562-s (CRL-3343) and WT SV40-immortalized mouse embryonic fibroblast (iMEF, CRL-2907) cells were purchased from ATCC, at which point they were certified mycoplasma-free. MDA-MB-231-L2 cells were a gift from Olga Razorenova, PhD (University of California, Irvine); they were isolated from lung metastases of MDA-MB-231 cells xenografted into murine fat pads. Cells were not authenticated or tested for mycoplasma contamination in our laboratory. HEK293T, U-2 OS, MDA-MB-231-L2, and iMEF cells were grown in DMEM, high glucose, GlutaMAX™ Supplement (Gibco #10566024), supplemented with 10% FBS (Sigma #12306C). K562-s cells were grown in RPMI 1640 medium (Gibco #11875093) supplemented with 10% FBS. All cells were grown at 37°C and 5% CO_2.

HEK293T cells were transfected with Fugene (Promega #E2311), using a ratio of 1 μg DNA to 3 μL Fugene mixed together in serum-free DMEM or OptiMEM (Gibco #31985062). Unless otherwise noted, the target site for mutation was the genomic HEK293 site3. All other cell types were transfected with Lipofectamine 3000 (Invitrogen #L3000015), using a ratio of 1 μg DNA to 1 μL P3000 reagent to 4μL Lipo3000 reagent, mixed together in OptiMEM.

To create the 293T-peCHYRON$_{6xHis}$ cell lines, HEK293T cells were transfected with plasmids expressing PE2 and a pegRNA that directs the installation of a 20-bp insertion based on the 6xHis sequence, then single colonies were isolated in two rounds of colony picking, dilution of isolated cells, and plating. Integration into the targeted genomic locus and how many of the three copies of this chromosome were modified were verified by deep sequencing. Which clone of these cell lines was used in each experiment is noted in Supplementary Table 4.3.

For transient transfections in 6-well dishes, 1.5 μg prime editor-expressing plasmid was mixed with 500 ng pegRNAs. To create cell lines using PiggyBac integration, cargo mixes were mixed with 1/10 – 1/20 total plasmid mass of Super PiggyBac Transposase Expression Vector (System Bioscience, Inc. #PB210PA-1). Stable integrants were selected with 1-2 μg/mL puromycin, 10 μg/mL blasticidin, or 700 μg/mL hygromycin. For the chemical induction experiments shown in Figure 4.5, cells were treated with 2 μg/mL doxycycline or 2 mM IPTG where indicated.

4.2.3 Screening pegRNAs for efficient prime editing

Unless stated otherwise, experiments toward finding efficient parts for peCHYRON utilized transient transfections of constructs encoding the indicated components with 3 days incubation before ending the experiment by freezing cell pellets. This pertains to Figure 4.2a-c, Supplementary Figure 4.1a-b, Supplementary Figure 4.1d-e, and Supplementary Figure 4.2a. For the experiment shown in Supplementary Figure 4.2b, plasmids encoding PE2 and a

pegRNA installing each B sequence were transfected into $293T_{6xHis}$ cells in two rounds over 8 days. Then, a plasmid encoding PE2 and libraries encoding the corresponding 6xHis-inserting pegRNAs were transfected into each sample, and cells were collected and analyzed after 2 days.

4.2.4 Lineage tracing assays

For the reconstruction shown in Figure 4.4 and Supplementary Figure 4.5, 293T-peCHYRON$_{6xHis}$ cells in a 6-well dish were transfected with 1.5 μg pCMV-PE2 and 0.5 μg of a mix of equal amounts of 42 pegRNA-expression vectors. These cells were allowed to grow for 8 days, passaging once. Then ("day one"), 20,000 of these cells were plated in each of 4 wells of a 96-well plate, then transfected the next day with the same mix of plasmids. On day four, when they had expanded to approximately 135,000 cells per well, the cells were trypsinized, and each well was split into 2 wells of a 24-well dish, before re-transfecting on day five. On day seven, the cells were trypsinized, and the entire contents of each well were moved to one well of a 12-well dish. On day nine, when cells had expanded to approximately 750,000 cells per well, each well was split into two wells of a 6-well dish. On day eleven, when each well had expanded to approximately 1.6 million cells, all wells were collected and analyzed by amplicon sequencing.

4.2.5 Illumina sequencing library preparation

Genomic DNA was isolated with a QIAamp DNA Mini Kit (Qiagen #51304). After DNA extraction, the appropriate region was amplified by PCR. The primers contained partial Illumina adapters and a 5 – 7 nt sample-specific barcode. The PCR was performed with Phusion HotStart Flex DNA Polymerase with GC buffer, with or without DMSO (NEB), 30-35 cycles.

For the experiments shown in Figure 4.2a-b; Supplementary Figure 4.1a and d; and Supplementary Figure 4.2c, reactions were pooled and purified by binding to AMPure beads

(0.9 beads:1 sample). The libraries were sent to Genewiz, Inc. for Amplicon-EZ sequencing, where they were further amplified to incorporate the TruSeq HT i5 and i7 adaptors and then sequenced on an Illumina HiSeq 2500 with a paired-end 250 protocol.

For the experiments shown in Figure 4.2c, Supplementary Figure 4.1e, Supplementary Figure 4.2a-b, Supplementary Figure 4.3a, Figure 4.3e, and Figure 4.4, reactions were purified individually by binding to AMPure beads (0.9 beads:1 sample). Then, they were further amplified to incorporate the TruSeq HT i5 and i7 adaptors, using Q5 High Fidelity DNA Polymerase with GC enhancer, for 15 cycles. The amplified libraries were pooled and purified again by binding to AMPure XP (Beckman-Coulter #A63880) beads (0.9 beads:1 sample). They were subsequently sequenced on an Illumina HiSeq using a paired-end 150-bp protocol at Novogene, Inc. Experiments shown in Figure 4.5 were prepared as above, except they were sequenced on an Illumina NovaSeq X using a paired-end 150-bp protocol at Novogene, Inc. Experiments shown in Supplementary Figure 4.6 were prepared as above, except they were sequenced on an Illumina MiSeq using a paired-end 300-bp protocol by the UCI Genomics Research and Technology Hub.

4.2.6 Nanopore sequencing library preparation

For the experiments shown in Supplementary Figure 4.1b, Figure 4.3a-d, Figure 4.3f, Supplementary Figure 4.5a-b, and Supplementary Figure 4.7, genomic DNA was isolated with a QIAamp DNA Mini Kit (Qiagen #51304). After DNA extraction, the appropriate region was amplified by PCR. The primers contained a 10-nt sample-specific barcode for multiplexing. The PCRs were performed with Phusion HotStart Flex DNA Polymerase with GC buffer, with or without DMSO (NEB), 30-35 cycles. The sequences then underwent end repair, A-tailing, and ligation of Oxford Nanopore adaptors using Oxford Nanopore Ligation Kit 14 per the manufacturer’s instructions, except all reagent volumes were divided in half. Libraries were then

sequenced on an appropriate MinION flow cell (R10.4.1) and basecalled using Oxford Nanopore's standard software.

4.2.7 High throughput sequencing analysis

The sequences retrieved by HTS were first demultiplexed based on their barcodes; for Illumina sequencing, the script is available at github.com/liusynevolab/CHYRON-NGS. For nanopore sequencing, sequences were demultiplexed using the Maple pipeline, available at https://github.com/gordonrix/maple. Maple outputs demultiplexed .bam files, which we converted to .fastq using the bedtools command bamtofastq. For the experiments shown in Supplementary Figure 4.1e and Supplementary Figure 4.2a-b, data were analyzed as in[14]. Scripts are available at github.com/liusynevolab/CHYRON-NGS. For the experiments shown in Figure 4.2a-b; Supplementary Figure 4.1a and d; BE samples in Supplementary Figure 4.1b, and Supplementary Figure 4.2c, data was analyzed with CRISPResso2[34]. For the high-throughput pegRNA screen shown in Figure 4.2c, 20-bp insertion sequences were extracted from fastq files by simply grabbing the 20-letter sequence at the expected edit site, then comparing it to the wt (unedited) sequence. If the extracted sequence exceeded a Hamming Distance cutoff from the unedited sequence, it was considered a real insertion for further analysis. All insertion sequences were tabulated, and instances of identical insertions were tallied. To calculate enrichment scores, Illumina sequencing reads from the pegRNA hp-miniprep pool (the plasmids that were transfected into $293T_{6xHis}$ cells at the start of the experiment) were analyzed in the same manner. The abundance of each RT template sequence in the original pegRNA pool was tabulated. Finally, the enrichment of each 20-bp insertion sequence was calculated as follows, where *g* is the proportion of genomic reads bearing that insertion sequence, and *m* is the proportion of library plasmid (the hp-miniprep pool) reads bearing that insertion sequence:

(1) $$log_2\left(\frac{g}{m}\right)$$

For Supplementary Figure 4.3a, Figure 4.3e, and Figure 4.4, for each sample, associated forward and reverse reads were merged (PEAR 0.9.10[35]). For Figure 4.5 and Supplementary Figure 4.6, forward reads were used. For Supplementary Figure 4.1b, Figure 4.3a-d, Figure 4.3f, Supplementary Figure 4.4a-b, and Supplementary Figure 4.7, nanopore sequences were used. Insertion sequences were extracted by finding the expected sequence motif upstream of the edit site, then searching for the expected sequence motif downstream of the edit site. For each read, the insertion length was increased in increments of 20 bp until the expected sequence motif downstream of the edit site was found. The full insertion sequence was extracted and the 17-bp propagator sequences in each insertion were compared to the known propagator sequences to ensure only legitimate insertions were grabbed. The 3-bp signature mutations in each insertion were converted to 1-character symbols to improve the ease of interpreting results. Insertion sequences represented by 1-character signature mutations were tabulated, and the counts of each insertion sequence were tallied. In parallel, the lengths of insertion sequences between the upstream motif and downstream motif were tallied, even if they were not multiples of 20 and/or did not contain the expected propagator sequences.

All the analyses were done in Python. The scripts and detailed instructions are available at github.com/liusynevolab/peCHYRON unless otherwise specified.

4.2.8 Calculation of cannibalization

For Figure 4.3a and Supplementary Figure 4.4a, nanopore sequencing of pegRNA genes was used to characterize gene length over time – the longer read lengths of which nanopore is capable allowed us to use amplicons >1 kb whose efficiency would not be biased by the insertions or deletions we are trying to detect. Because nanopore sequencing produces a significant rate of insertion and deletion errors, pegRNA gene sequences were compared to sequences of the original pegRNA gene plasmid, amplified under the same conditions at the same time and sequenced in the same run. After demultiplexing as described above, the sizes

of all sequences that lie between a motif upstream of the pegRNA coding sequence and a motif downstream of the coding sequence and protective nick site was computed. The range of sizes that included 98% of reads for the plasmid was defined as the “unchanged” size. Then the proportion of sequences that were shorter or longer than the unchanged size was computed for each sample, and the proportion of such sequences in the plasmid sequence was subtracted from the value.

4.2.9 Calculating editing efficiencies and modeling the expected distribution of edits from a long timecourse

To determine if the rate of editing at the peCHYRON locus remains constant as propagation proceeds, experimental samples from a 67-day timecourse were analyzed in two different ways—for one, we looked at editing rates as a function of time. Secondly, we characterized editing rates as a function of recording locus length. For the assessment of editing rates over time, the apparent efficiencies of A→B and B→A pegRNAs (% editing / day) were calculated from samples collected at various timepoints throughout the 67-day experiment (Figure 4.3c). In an ideal case where editing efficiency stays constant, the A→B and B→A efficiencies would be expected to stay the same at each time point. To characterize the relationship between editing rate and recording locus length, the apparent A→B and B→A efficiencies at the end of the 67-day timecourse were used to calculate the expected distribution of insertion lengths for the ideal case where editing efficiency is unaffected by the recording locus’s length. The distribution of edits from this ideal case was then compared to our experimental data (Figure 4.3d).

For both of these analyses, peCHYRON’s editing dynamics were modeled as a system of first order rate laws that describe the rate of converting recording loci from having n edits to n+1 edits. Separate rate equations were written for loci bearing different numbers of edits, and two different efficiency values were implemented to reflect the different efficiencies of A→B and

B→A pegRNAs. Example rate equations describing the conversion from wt to singly edited loci, singly to doubly edited loci, and doubly to triply edited loci are shown below.

(2) $$\frac{d[wt\ targets]}{dt} = -\eta_{A\to B} * [wt\ targets]$$

(3) $$\frac{d[1\ edit]}{dt} = \eta_{A\to B} * [wt\ targets] - \eta_{B\to A} * [1\ edit]$$

(4) $$\frac{d[2\ edits]}{dt} = \eta_{B\to A} * [1\ edit] - \eta_{A\to B} * [2\ edits]$$

where t is in units of days, $\eta_{A\to B}$ and $\eta_{B\to A}$ denote the per-day editing efficiencies of A→B and B→A pegRNAs, respectively, and brackets denote concentrations (*i.e.*, percentage of loci with the specified number of edits). [wt targets] indicates the fraction of wildtype, viable peCHYRON recording loci, where "viable" means they are capable of being edited. In experiments with non-clonal populations of cells, we and others with a similar prime editing-based recorder[15] have consistently observed a population of cells that are not amenable to editing, while other cells acquire edits efficiently. This is presumably due to cell-to-cell variability during transformation of transgenes, cell cycle differences, and other unknown factors. Thus, to get an accurate calculation of the efficiency of A→B edits, it is important to not model the entire population of wildtype recording loci as the substrate for the first edit, but instead to model only viable wildtype loci.

To solve for $\eta_{A\to B}$ and $\eta_{B\to A}$, equations (2) through (4) were integrated to yield the following:

(5) $$[wt\ targets](t) = e^{-\eta_{A\to B}*t} * [wt\ targets]_0$$

(6) $$[1\ edit](t) = \frac{e^{-\eta_{A\to B}*t-\eta_{B\to A}*t}*\left(e^{\eta_{A\to B}*t}-e^{\eta_{B\to A}*t}\right)*\eta_{A\to B}*[wt\ targets]_0}{\eta_{A\to B}-\eta_{B\to A}}$$

(7) $$[2\ edits](t) = \frac{e^{-\eta_{A\to B}*t-\eta_{B\to A}*t}*\eta_{A\to B}*\eta_{B\to A}\left(e^{\eta_{A\to B}*t}-e^{\eta_{B\to A}*t}-e^{\eta_{B\to A}*t}*\eta_{A\to B}*t+e^{\eta_{B\to A}*t}\eta_{B\to A}*t\right)[wt\ targets]_0}{(\eta_{A\to B}-\eta_{B\to A})^2}$$

Where [wt targets]$_0$ represents the initial fraction of viable peCHYRON loci. To assign a numerical value to [wt targets]$_0$, we used empirical data from the 67-day timecourse. Specifically, we observed that the fraction of edited loci plateaued at ~82% after 31 days, despite continuous accumulation of long insertions. In other words, the cells that contained viable target sites kept propagating, while unedited targets in the population seemed to stagnate. This led us to believe the editing machinery remained functional over time, but the targets that never acquired an edit were simply not viable targets. Thus, the maximum fraction of recording loci that acquire 1 or more edits by the end of the long timecourse is approximately equal to [wt targets]$_0$.

Because we assume all viable targets acquire one or more edits during the long timecourse, equation (5) cannot be used to solve for $\eta_{A\to B}$ at all timepoints (efficiency approaches infinity as the fraction of viable wt targets approaches 0). Thus, equations (6) and (7) were instead used to solve for $\eta_{A\to B}$ and $\eta_{B\to A}$. They were rearranged to their root forms (*i.e.*, they were rearranged to equal 0), then the Scipy root finder (scipy.optimize.root) was used to find the values of $\eta_{A\to B}$ and $\eta_{B\to A}$. The python script that implements the Scipy root finder can be found at github.com/liusynevolab/peCHYRON.

To calculate the expected distribution of insertion lengths for an ideal recorder that does not lose editing efficiency as loci grow longer, rate equations were written for loci bearing all insertion lengths up to 6 edits, following the same pattern as equations (2)-(3) above. The resulting system of ordinary differential equations (ODEs) was solved in Mathematica at t = 67 days using the $\eta_{A\to B}$ and $\eta_{B\to A}$ values calculated from experimental data, yielding the expected

fraction of each insertion length. The Mathematica notebook with this system of ODEs is available at github.com/liusynevolab/peCHYRON.

4.2.10 Calculation of corruption

We integrated all the necessary components of peCHYRON into the genome in 293T cells, including synthetic recording loci. Cells were transfected with a mix of PiggyBac cargo including synthetic recording loci, PEmax, protective nicking guide, and small libraries of A→B (10 different signatures present in the library) and B→A pegRNAs (21 different signatures). 8 populations each founded by 1,000 cells, then split into 4 replicates each for outgrowth and sequencing, were used to better characterize peCHYRON recording. We used this dataset to determine the rate of corruption. To avoid undercounting corrupted loci, we stringently defined a "correct read" as an Illumina HTS read that contained either no edit or an insertion whose length is an exact multiple of 20. A "corrupted read" was one that contained any insertion that is not a multiple of 20, or any deletion. As an unbiased approximation of when the corruption occurred, we binned each corrupted sequence with the round of recording corresponding to the closest size. Seven of eight populations showed a low rate of corruption at each step. In the eighth population, we observed an unusual type of corruption. Some loci in each of the four replicates of this population acquired insertions of 48 bp, 73 bp, or 78 bp that are repeats of all or part of the synthetic recording locus. Then, these aberrant insertions acquired additional genuine 20-bp insertions. These genuine insertions made it appear that there were large numbers of, for example, 88-bp insertions (48 bp+20 bp+20 bp), creating a false impression that the majority of loci with 4-5 rounds of recording were corrupted. We properly assigned the aberrant insertions as what they were—errors in the first round of recording—and removed the further genuine insertions from the "corrupted" list.

4.2.11 Lineage reconstruction

To reconstruct cell lineage in Figure 4.4, we created a list of all insertion sequences in each of the 13 wells used for the analysis. Each insertion has an abundance, based on the number of HTS reads that include that exact insertion sequence, and a length, equal to the number of peCHYRON insertions at the recording site. For our initial analysis, the researcher performing the analysis (TBL) was not told which well was which. We refined the list for each well to include only those insertions that passed two filtering steps: 1) a length filter that excluded any sequences with 3 or fewer inserted sequences and 2) an abundance filter that removed everything after a decreasing convex knee found using the kneedle algorithm[26]. For each sequence in this set, we generated all possible subsequences containing the first insertion (*i.e.* prefixes) and assigned a weight equal to the inverse of the number of prefixes. This results in a multiset of prefixes for each sequence. Afterwards, the distances between the multisets were found using the generalized Jaccard distance for multisets. We used this set of distances to reconstruct the relationships using the UPGMA[36] hierarchical clustering algorithm. (github.com/scipy/scipy/blob/v1.2.1/scipy/cluster/hierarchy.py#L411-L490). The two algorithms (Jaccard and Prefix Jaccard) were compared by computing the Robinson Foulds score[37], a metric that compares a reconstruction to a ground-truth tree, for each reconstruction (Supplementary Figure 4.5b). For each percentage of downsampling between 99 and 5%, 1000 different data downsampling operations were performed. Robinson Foulds scores were computed for each using the ete3 package[38] (http://etetoolkit.org/).

All the analyses were done in Python. The scripts and detailed instructions are available at github.com/liusynevolab/peCHYRON.

4.2.12 Event reconstruction *via* bigram frequencies

To evaluate bigram frequencies of signatures at recording loci, ‘expanded bigrams’ were defined as any ordered pair of signatures in a record, taking into account non-contiguous pairs.

In other words, a single signature produces a bigram for each signature after it in a record (*e.g.*, for a four-signature record ABCD, the expanded bigrams produced from signature A are AB, AC, and AD; the expanded bigrams of signature B are BC and BD). The read counts of all bigrams were enumerated by sample to assess the total frequency of each bigram across all insertions. For transfection epoch experiments in Supplementary Figure 4.6b, after removal of signature sequences not included in the overall experiment (the result of sequencing error), the total bigram frequencies were plotted as a heatmap with the signature in position 1 ('P1') and position 2 ('P2'). For chemical epoch experiments in Figure 4.5b, total bigram frequencies were divided by frequencies of the bigram's anadrome (*e.g.,* doxycycline-IPTG frequency divided by IPTG-doxycycline frequency), and then $\log_2$ transformed to obtain the $\log_2$ of the fraction of a bigram vs. its anadrome.

4.2.13 Event reconstruction *via* stretched records

Samples were sequenced with Illumina. The resulting fastq files were aligned to a reference sequence for the peCHYRON recording locus, and insertions at the recording locus were extracted. The insertions were analyzed in 20 bp increments, where the first 3 bp are the signature mutation and the last 17 bp are the propagator sequence. If propagator sequences didn't match their expected sequences within a hamming distance of 2, they were assumed to be incorrect amplicons from sequencing library prep and were thrown out. If the insertions passed this first filter, the 3 bp signature mutations were extracted and converted to 1-letter symbols according to a user-defined signature key (this is to simplify interpretability of records). Then, the resulting records were filtered again to remove reads with unexpected signatures due to sequencing error (for example, samples from the dox and IPTG recording experiment should only have signatures corresponding to the dox- or IPTG-inducible pegRNAs or the constitutive pegRNA). This final list of records was computationally "stretched" to the least common multiple of the lengths of all records. Insertion lengths with fewer than 100 reads were excluded from the

analysis. To assign weights to records of different lengths, the number of reads corresponding to each insertion length was first tallied. Then each insertion length was given a weight as described in equation 10.

(8) $$weight\ for\ a\ given\ insertion\ length = \frac{n\ reads\ of\ the\ most\ abundant\ insertion\ length}{n\ reads\ of\ the\ insertion\ length\ of\ interest}$$

For example, if records with 2 insertions were most abundant, and there were 100,000 records with 2 insertions but only 20,000 records with 3 insertions, the weight for records with 3 insertions would be 5 (100,000 / 20,000). When calculating the fraction of each signature at each insertion length, records were pooled together according to their weight (*e.g.*, a record given a weight of 5 was added to the pool 5 times). The fraction of each signature is simply the % of records in the pool with a particular signature at each position.

For the samples exposed to chemical epochs (Figure 4.5c), instead of plotting the fraction of each signature at each position, the fold-change relative to the uninduced case was plotted. The fraction of each signature in the uninduced case was calculated from a control sample that was never exposed to any inducer for 18 days. To calculate fold-change, the experimental samples' fractions of each signature at each position were first computed, then divided by the fractions observed in the uninduced control sample. To calculate pairwise Manhattan distances between the stretched records of pairs of samples, a 120-dimensional vector was created for each sample, consisting of the fold changes at each of the 60 points on the stretched array, for both doxycycline-inducible and IPTG-inducible signatures. All pairwise Manhattan distances were then computed.

All these analyses were conducted in Python. Refer to github.com/liusynevolab/peCHYRON for scripts and detailed instructions for use.

4.2.14 Entropy calculations

To calculate Shannon entropy[25], we first made a table of all the signature sequences in the relevant dataset with, for each sequence, the number of times it was observed (the "count"). We did one calculation for signatures inserted at "odd" positions (1, 3, 5, etc.) and one for "even" signatures. Once a table of signatures and counts (c) was created, we calculated the proportion (*p*) for each sequence (equation shown here for sequence i).

(9) $$p_i = \frac{c_i}{c_1 + \cdots + c_i + \cdots + c_n}$$

Then we calculated the overall Shannon entropy (H) for the dataset:

(10) $$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

All analyses were done in Excel.

4.2.15 Statistical analyses

In all cases, biological replicates were derived from different populations of cells that were manipulated separately throughout the experiment. For technical replicates, cells were grown and manipulated, and DNA extracted, together. All procedures downstream of DNA extraction were performed separately. Significance of the results in Supplementary Figure 4.7 was calculated in GraphPad Prism by Brown-Forsythe ANOVA test, because 3 of the groups failed the Shapiro-Wilk test of normality.

4.2.16 Figure preparation

Figure 4.1, Supplementary Figure 4.1c, Supplementary Figure 4.3b, Figure 4.5a, and Supplementary Figure 4.7a were prepared with InkScape. Figure 4.2c, Figure 4.5c, and Supplementary Figure 4.6c were plotted in Excel, then converted to scalable vector graphics with InkScape. A portion of Figure 4.4a was prepared at biorender.com. The plots in Figure 4.4b-c and Supplementary Figure 4.5a were generated using the hierarchy.dendrogram function in matplotlib (scipy.org). All other plots were created in GraphPad Prism. Heatmaps of transfection epoch bigram frequencies in Supplementary Figure 4.6b were plotted using the heatmap() function of the seaborn library in Python3, and bar graphs of $\log_2$(doxycycline-IPTG/IPTG-doxycycline) frequency fractions in Figure 5b were plotted with the barplot() and stripplot() functions of the seaborn library in Python3.

4.3 Results

4.3.1 Architecture of peCHYRON

Our goal with peCHYRON was to build a recorder that couples high durability and extensibility with the high information content that is easily accomplished with prime editor. To achieve high durability, we used the PE2 instantiation of prime editor, in which only one DNA strand is nicked, so the unintended mutation rate is extremely low (as little as 1% as frequent as the intended edit)[21]. To achieve high extensibility, we designed peCHYRON so that each round of recording installs a target site for the next round. First, a 20-bp locus in a mammalian cell's genome is targeted by a pegRNA. The pegRNA directs prime editor (PE2max)[21,23] to reverse transcribe the pegRNA's programmable template sequence into the locus. The reverse transcriptase (RT) template sequence is programmed to insert a variable 3-nt sequence that encodes up to 6 bits

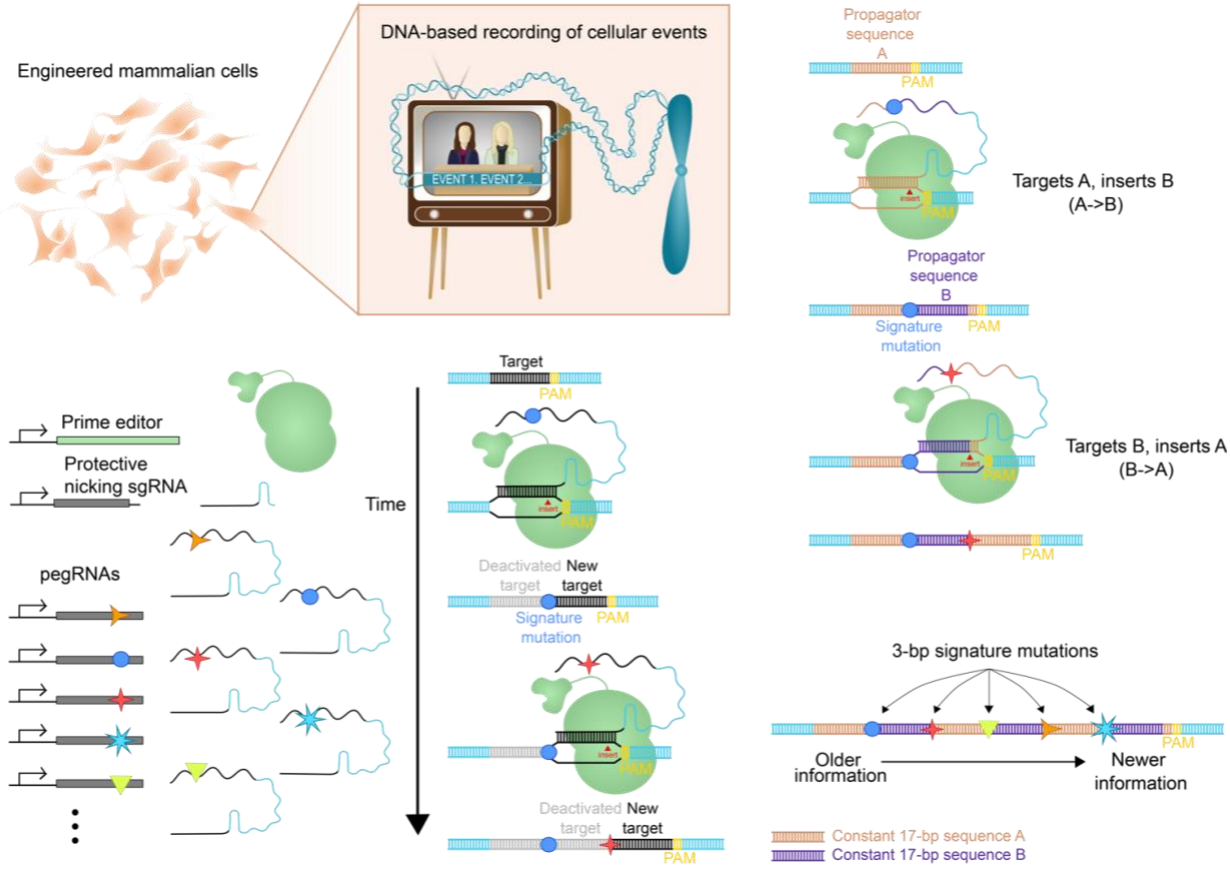


Figure 4.1. Overview of peCHYRON. **a,** Using peCHYRON, transient events are recorded into DNA in the order they occur, as in a scrolling chyron. **b,** Iterative sequential insertion of signature mutations *via* prime editing. Signature mutations are variable 3-bp sequences, each theoretically encoding up to 6 bits of information. A collection of pegRNAs facilitates the insertion of signature mutations at the recording locus during continuous and iterative editing. In a single cycle of prime editing, a signature mutation and a 17-bp propagator sequence are inserted at the recording locus adjacent to the PAM. The newly-inserted propagator sequence forms the target site for the next cycle of editing because it is now PAM-adjacent. The previous target site is deactivated due to separation from the PAM. This process can repeat indefinitely, resulting in ordered accumulation of regularly-spaced signature mutations. **c,** Propagation with alternating pegRNAs. To avoid issues with repetitive sequence elements, the recording locus alternates between addition of two different 17-bp propagator sequences. In one step, a pegRNA targets sequence A and inserts sequence B. In the next step, a pegRNA targets sequence B and inserts sequence A. The process repeats and continuous propagation of the recording locus proceeds without end. **d,** How to interpret information at a peCHYRON locus. Information-rich signature mutations are separated by constant 17-bp propagator sequences, resulting in clear punctuation between sequential editing events. The age of a signature is reflected by its position in the record, with PAM-proximal signatures reporting on the most recent events.

of information followed by a constant 17-nt propagator sequence that serves as the target sequence for the next insertion step (the 3 bp nearest to the PAM are not altered during an edit, so only 17 nts of the target need to be re-installed to constitute a new 20-bp target adjacent to the PAM). After one cycle of insertion, the previous propagator sequence is no longer PAM-adjacent, and therefore inactive, while the new propagator sequence that forms the new 20-bp target site becomes PAM-adjacent, and therefore active. The next step of editing inserts another variable 3-nt sequence along with the next propagator sequence, iterating indefinitely. In this manner, each variable 3-bp sequence, which we refer to as a signature mutation, records an event. Because the information content of the signature mutation can be up to 6 bits, peCHYRON has sufficient complexity to theoretically record 64 different types of events. New signature mutations appear in the order of events recorded. Constant 17-bp sequences separate signature mutations, acting as punctuation. The recording process does not change previous records, as edits occur through sequential insertions, and recording can, in principle, propagate continuously without end (Figure 4.1b).

The simplest implementation of peCHYRON would involve the sequential insertion of a constant propagator sequence (Supplementary Figure 4.1a-c). However, iterative addition of a single propagator sequence requires a pegRNA whose RT template sequence (containing the

17-nt propagator sequence to be inserted) and primer binding sequence (PBS)[21] share identity. During the RT step of prime editing, this would allow primer binding in the pegRNA's RT template sequence rather than the PBS, resulting in no net insertion (Supplementary Figure 4.1c). We therefore settled on the use of two alternating 17-bp propagator sequences in our peCHYRON design so that the RT template sequence and PBS in each pegRNA are distinct. The peCHYRON architecture we implemented is shown in Figure 4.1c where one pegRNA targets site A and adds a propagator sequence B (A->B) and the other pegRNA targets sequence B and adds sequence A (B->A), allowing for propagation. Importantly, although peCHYRON uses alternating propagator sequences, this design does not require 2 edits to record a single event. For any event that we wish to record, we simply express a pair of A->B and B->A pegRNAs simultaneously, so one of the pegRNAs corresponding to the event is compatible with editing the recording locus at any time. Recorded information in the resulting peCHYRON locus is easily interpretable (Figure 4.1d).

4.3.2 Identification of efficient parts for peCHYRON

In the original report of prime editing by Anzalone *et al.*[21], it was found that an 18-bp sequence ("6xHis") could be inserted at target genomic sites using PE3, a prime editor that achieves high genome-editing activity by nicking both DNA strands. Since the core mechanism of peCHYRON involves the insertion of a 20-bp sequence (*i.e.,* a 3-bp signature mutation and a 17-bp propagator sequence), PE3's ability to insert 6xHis was a natural starting point in the design of peCHYRON. Several developments were necessary.

First, we reasoned that the modest rates of unintended insertion or deletion mutations (indels) associated with PE3[21] could be problematic in peCHYRON, because spurious indels at any step of the iterative insertion process would terminate recording. We therefore tested whether 6xHis could be inserted using PE2, which does not cut both strands of target DNA[21] (Figure 4.2a). To our surprise, we found that 6xHis was inserted by PE2 at comparably high

efficiency as by PE3. As expected, the 6xHis insertion made by PE2 was accompanied by a much lower rate of unintended insertion or deletion mutations (indels). Consequently, we decided to use PE2 as the prime editor for peCHYRON.

Second, we tested whether the 18-bp 6xHis insertion sequence identified in Anzalone *et al.*[21] could be converted into a 20-bp insertion sequence in which the first 3 bps constituting the signature mutation are randomizable (Figure 4.2b). A collection of pegRNAs, each of which targets a defined site in the genome ("site 3"[21,24]) and contains the template for inserting a

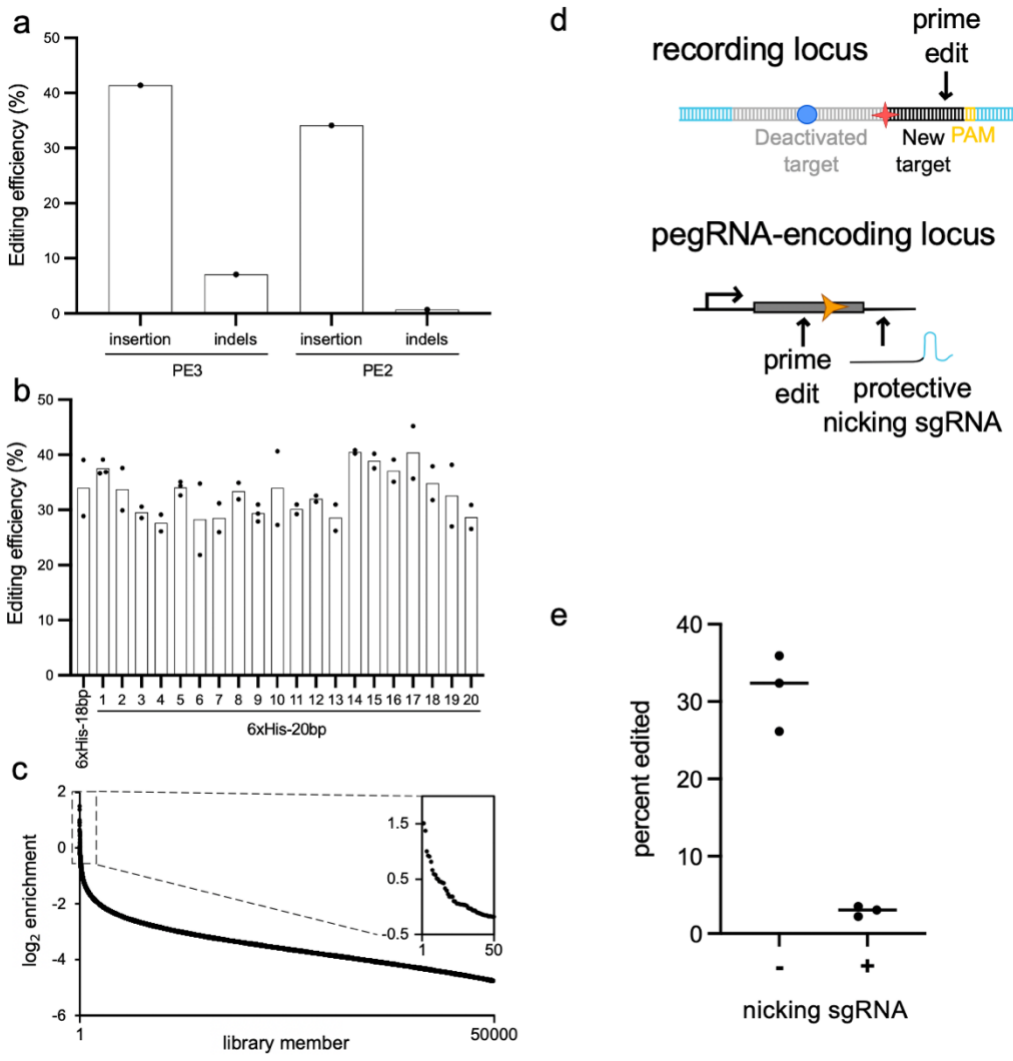


Figure 4.2. Identifying efficient components for iterative editing with two alternating propagator sequences. **a,** PE2, a prime editing system that does not make double-strand breaks, efficiently adds insertions at the peCHYRON locus. An 18-bp 6xHis sequence was inserted at site3 in HEK293T cells using PE3 and PE2, and the rates of intended insertions vs. spurious indels were assessed. **b,** 20-bp sequences derived from the 18-bp 6xHis sequence were inserted efficiently. Each 20-bp sequence was nearly identical to the original 6xHis sequence, but with the first 3 bp randomized (3-bp signature + 17-bp propagator sequence). Insertions were made at site3 in HEK293T cells using PE2. Points represent biological replicates. Efficiencies from different experiments were normalized using samples that were present in each experiment (6xHis-20bp sequences 1, 5, and 9). **c,** Efficient pegRNAs that target 6xHis and insert a different 20-bp sequence (A→B) were rare. A high-throughput screen was used to identify pegRNAs that could alternate with the 20-bp variant of the 6xHis tag to iteratively edit the peCHYRON locus. In the screen, a library of ~100,000 pegRNAs with variable insertion sequences was transfected into HEK293T cells along with PE2. High enrichment values correspond to pegRNAs that frequently edited the 6xHis target site without being over-represented in the original transfection mix. The top 50,000 enriched library members were plotted, and the inset shows the top 50. **d,** Cartoon showing a strategy for blocking prime editing in the genes that encode the pegRNAs. Top arrow indicates the site at which the peCHYRON recording locus is nicked by prime editor to initiate the next prime edit. Bottom arrows indicate the site at which the pegRNA genes could be cannibalized (see Supplementary Figure 4.3b) and the adjacent nick site on the same strand that could prevent this cannibalistic prime edit. The protective nick site is present in the pegRNA-expressing loci only, so the pegRNA genes are protected from prime editing, whereas the recording locus can be edited as desired. **e,** Prime editing at a genomic locus was blocked by nicking an adjacent site on the same strand. 293T cells were transiently transfected with plasmids encoding PEmax, a pegRNA that directs a 20-bp insertion at the HEKsite3 locus, and either an sgRNA that directs a nick at a distant locus (-) or an sgRNA that directs a nick 21 nt 5' of the cannibalization-associated nick (+), on the same strand. 3 days after transfection, the genomic site was sequenced. The graph indicates the proportion of loci that had been prime edited; each dot represents one biological replicate.

distinct 3-nt signature mutation followed by a constant 17-nt sequence nearly identical to 6xHis, was tested for PE2-mediated editing in HEK293T ("293T") cells. Although some variation was observed, most sequences were inserted with high efficiency similar to that of the original 18-nt 6xHis sequence.

Third, we attempted to make a pegRNA that uses the inserted 6xHis-based sequence as its target, to be used for subsequent propagation. Recall that the 6xHis-based sequence was inserted at the site3 target. If we consider site3 to be A and 6xHis to be B (see Figure 4.1c), the original pegRNA that installed 6xHis corresponds to the A->B step, and a pegRNA designed to insert site3 at the 6xHis target corresponds to the B->A step. We designed a B->A pegRNA that inserts site3. However, we did not observe efficient editing (Supplementary Figure 4.1d). To find a sequence that could be inserted efficiently, we performed a high-throughput screen in which a library of ~100,000 pegRNAs, each designed to target the 6xHis site and add a random 20-nt sequence, was transfected into 293T cells that contained the target 6xHis site in their genomes ($293T_{6xHis}$). After 3 days, we sequenced the target sites to identify enriched insertions, which we

then validated individually for high insertion activity. Efficient insertion sequences were uncommon, but several were identified (Figure 4.2c and Supplementary Figure 4.1e). For the ten most efficient insertion sequences, we subsequently optimized PBS length and ensured they could tolerate variable signature mutations (Supplementary Figure 4.2a).

For propagation to occur, an inserted sequence must act as the target for the next round of prime editing. Of the ten efficient insertion sequences, we identified two that tolerated variable signature mutations and behaved as good targets for the insertion of 6xHis sequences containing signature mutations (Supplementary Figure 4.2a-b). Let us call these two sequences from the high-throughput screen B_4 and B_7 and the 6xHis sequence A (see Figure 4.1c). The downselection from the screening steps described above ensured that B_4 and B_7 could successfully act with A both in the A->B step (*i.e.*, target A and insert B_4 or B_7) and the B->A step (*i.e.*, target B_4 or B_7 and insert A), providing the necessary ingredients for peCHYRON. When we simultaneously transfected A->B and B->A pegRNA pairs for both B_4 and B_7, we observed extended propagation (Supplementary Figure 4.2c). One pair resulted in slightly more efficient propagation (Supplementary Figure 4.2c), so we performed all subsequent experiments with that pair.

4.3.3 peCHYRON accumulates edits over time at genomic loci

The implementation of peCHYRON requires that all components be delivered to the relevant cells, integrated into their genomes, and maintained over time. Therefore, we created a $293T_{6xHis}$ cell line into which we have integrated all peCHYRON components: a synthetic target site comprised of the 20-bp 6xHis sequence inserted at the genomic site3 locus, prime editor, and A$\rightarrow$B and B$\rightarrow$A pegRNAs. Ideally, we would hope to see that progressively longer records accumulate at the recording locus, while the genes encoding the recording components remain stable. We observed the accumulation of edits at the recording locus as expected; however, we

also observed insertion mutations in the genes encoding the pegRNAs (Supplementary Figure 4.3a). This problem is inherent in the sequence constraints of peCHYRON components – because each pegRNA installs a target site to be used in the next round of prime editing, the sequence of that target site must also be present in the genes that encode the pegRNAs, themselves; therefore, A→B pegRNAs can target the B→A pegRNA genes, and vice versa (Supplementary Figure 4.3b). Left unchecked, this "cannibalization" by peCHYRON would destroy the pegRNA genes at the same rate that desired editing occurs at the recording locus. Truly long-term recording would be impossible, as editing efficiency would gradually fall due to depletion of intact pegRNA genes. To unlock peCHYRON's long-term functionality, we designed a method to block prime editing at the pegRNA genes without also blocking the desirable editing at the recording locus. Anzalone et al.[21] showed that, after installing a prime edit, nicking the opposite DNA strand biases host cell DNA repair to maintain the prime edit (PE3). We reasoned that if, after installing a prime edit, we instead nicked the same strand, we would get the opposite result: the host cell would reverse the prime edit. We call this idea "protective nicking." By targeting protective nicking to a sequence adjacent to pegRNA genes but not present at the recording locus, cannibalization could be selectively blocked without impairing the efficiency of the desired recording (Figure 4.2d). To test this idea, we first attempted to block prime editing *via* protective nicking at an arbitrary genomic site. We transiently transfected 293T cells with plasmids encoding PEmax and a pegRNA directing a 20-bp insertion at the genomic site. To implement protective nicking, we co-transfected a plasmid encoding an sgRNA that would cause nicking near the prime edit, on the same strand (21 nt 5' of the prime edit). As a negative control, we instead transfected a plasmid encoding an sgRNA that directs a nick at a distant genomic site. When the protective nick was made adjacent to the prime edit, prime editing was reduced ~10-fold (Figure 4.2e).

We next used protective nicking to prevent cannibalization, achieving a stable peCHYRON cell line that was able to record at a constant rate for more than two months. To

create this stable cell line, we integrated 3 PiggyBac cargo plasmids into $293T_{6xHis}$ cells: one encoding a protective nicking sgRNA and PEmax, marked with puromycin resistance; a second encoding an A→B pegRNA; and a third encoding a B→A pegRNA. Both pegRNA plasmids include a target site for the protective nicking guide and are marked with mCherry. An ideal cell line would exhibit efficient recording at the intended locus but low cannibalization. To achieve this behavior, we optimized the relative concentrations of PE, protective nicking guide, and pegRNAs by transfecting different relative amounts of their cargo plasmids along with PiggyBAC transposase. The cell line with the best combination of parameters at 18 days had a 2:1:1 molar ratio of PE/nicking guide to A→B pegRNA to B→A pegRNA plasmids and was first selected with puromycin and then sorted for high (top-20%) mCherry fluorescence to acquire cells with high expression of pegRNAs. This cell line was passaged for a total of 67 days post-transfection, along with a control line that was identical except that it did not express a protective nicking guide. In both samples, most cells recorded at least one edit at the recording locus (Figure 4.3a). The control cell line without protective nicking accumulated edits in the B→A pegRNA gene at the same rate as at the recording locus (Figure 4.3a). This is expected, because both the first edit at the recording locus and B→A cannibalization are directed by A→B pegRNAs. In contrast, in the cell line expressing the protective nicking guide, most B→A pegRNA genes remained intact throughout the experiment (Figure 4.3a and Supplementary Figure 4.4a). Cannibalization of A→B pegRNA genes was low with and without protective nicking (Figure 4.3a). Overall, the protected cell line exhibited extensive editing at the recording locus, with edits continuously accumulating for at least 67 days (Figure 4.3b and Supplementary Figure 4.4b).

The rate of editing remained constant both over time (Figure 4.3c) and over rounds of editing (Figure 4.3d). We investigated the rate of editing over time by collecting samples at various timepoints throughout the 67-day timecourse, then calculating the rate of editing for A→B and B→A steps (% editing / day) based on the observed proportion of loci with 0, 1, or 2 edits at each timepoint (see Methods). The rates were approximately constant over the entire

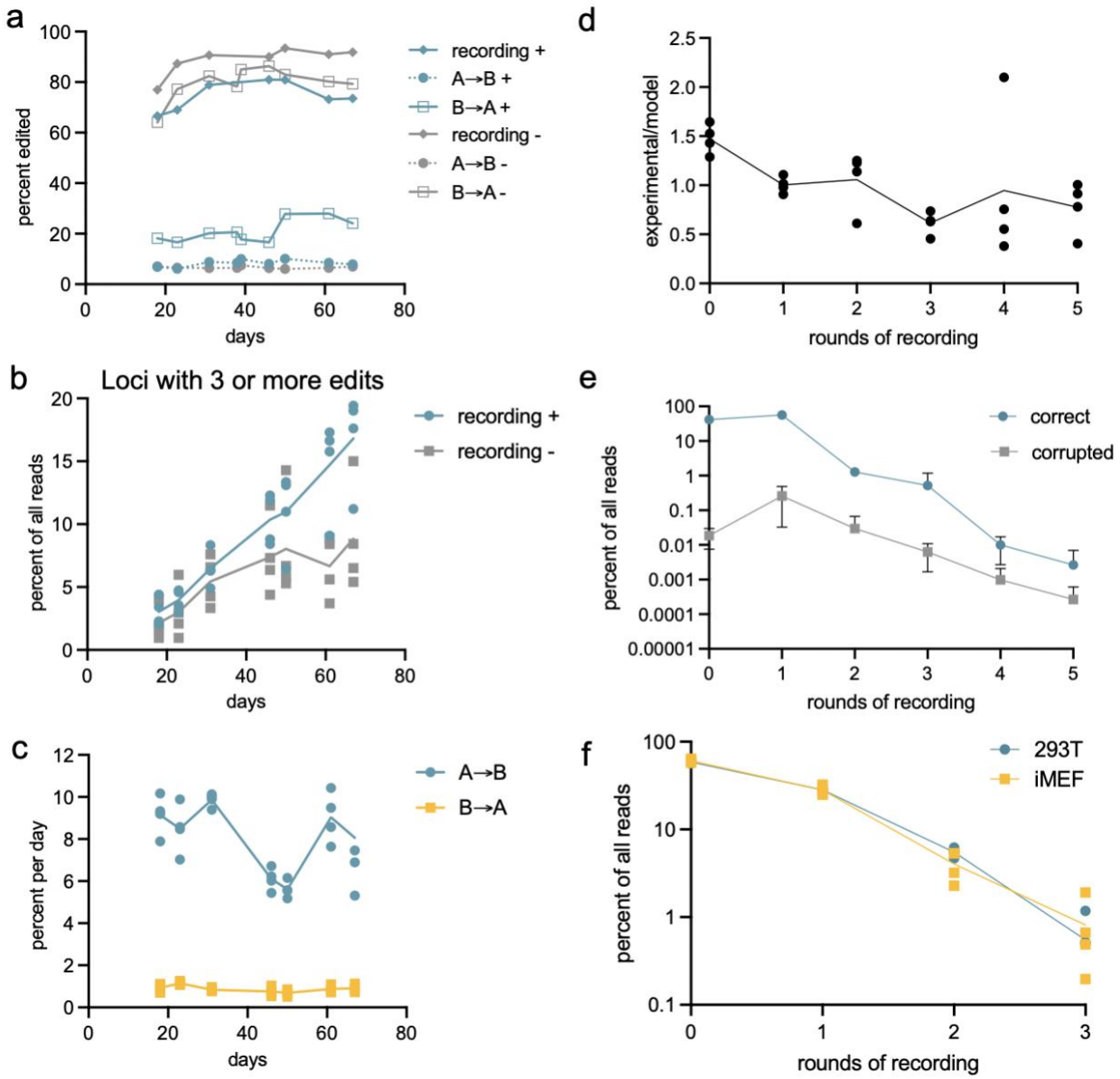


Figure 4.3. Characterization of peCHYRON performance. **a,** Cannibalization of the peCHYRON system was blocked by a protective nick. Expression cassettes for PEmax, A→B and B→A pegRNAs, and, optionally, protective nicking guides, were co-transfected with PiggyBac transposase for stable integration into the genome of $293T_{6xHis}$. Cells were grown for the indicated time after the initial transfection, then the recording locus, the A→B pegRNA-expressing loci, and the B→A pegRNA-expressing loci were sequenced. For the recording locus, dots represent the proportion of reads in which one or more of the expected peCHYRON edits were observed. + indicates samples with the protective nicking guide, - represents those without. Dots are the mean of 4 technical replicates. For the pegRNA-expressing loci, any deviation from the original sequence was considered cannibalization and plotted (see Methods). **b,** In the experiment shown in (a), at the recording loci, insertions accumulated for at least 67 days. Insertions of 20 bp were considered to correspond to 1 round of insertion, 40 bp to 2 rounds, and so forth. The fraction of insertions that were 60 bp or larger are plotted. Each point represents one technical replicate. 4 replicate values are shown for each round; some points overlap. **c,** Editing rates did not diminish over time. The efficiency of peCHYRON recording was computed at each timepoint. Each point represents one technical replicate. 4 replicate values are shown for each round; some points overlap. **d,** Editing rates did not diminish throughout rounds of sequential insertion. The efficiency values (A→B and B→A) calculated from the 67-day timepoint were used to calculate the distribution of insertion lengths that would be expected if the probability of each locus gaining an edit was independent of the locus's past edits. Experimentally observed distribution of insertion lengths at the 67-day timepoint was normalized to the expected distribution, so a value of 1 indicates that the observed efficiency matched the expected efficiency. Each point represents one technical replicate. 4 replicate values are shown for each round; some points overlap. **e,** The rate of spurious indel mutations during 24 days of passaging was very low. From a deeply-sequenced experiment in which 21 different pegRNAs are expressed, each read was assigned to a bin according to its size; insertions that exactly match the expected sizes were considered "correct," and all other mutations were considered "indels." The results of 32 biological replicates, representing 8 populations, are plotted. Error bars correspond to the 95% confidence interval. **f,** peCHYRON recording was similarly efficient in another cell type. peCHYRON recording cell lines were created from 293T and SV40-immortalized mouse embryonic fibroblasts in parallel. Recording loci were sequenced 15 days after transfection. Two technical replicates for each of two biological replicates are shown.

timecourse (Figure 4.3c), indicating peCHYRON can facilitate very long-term recording. We note that the rate of editing was much higher for A→B than B→A steps (Figure 4.3c). This is because the A→B step is driven by a pegRNA that was enriched in a high-throughput screen, whereas the B→A step simply inserts a modified version of the 6xHis sequence that was previously shown to insert efficiently at an endogenous site3 locus. In other words, the pegRNA that carries out B→A insertions has not yet been extensively optimized. The efficiency of the system should be greatly improved in the future if optimization of insertion sequences can bring the rate of B→A editing to the level currently achieved by A→B.

Aside from maintaining constant editing rates over time, long-term recording also requires that editing remains efficient as the recording locus accumulates edits (*i.e.*, the target site must not lose its receptiveness to editing as a result of past edits). To test whether this is the case for peCHYRON, we used the A→B and B→A editing rates calculated at the 67-day timepoint to model the expected distribution of edits that would be observed if the efficiency of

each edit were independent of past edits. The predicted distribution of edits closely matched our experimental data (Figure 4.3d), suggesting that editing does not diminish as loci grow longer. In contrast, for recorders that rely on DNA double-strand breaks, efficiency declines quickly for each successive edit at the same recording locus[10,11,13,14,18].

We next considered another parameter relevant to long-term recording: "corruption," or how often a round of recording led to an unintended edit at the recording locus. Because unintended edits often terminate recording, and corruption would progressively accumulate over rounds of recording, even a moderately high rate (*e.g.*, 10%) of corruption would be incompatible with long-term recording. We created a stable peCHYRON cell line, then measured corruption after 24 days of outgrowth in 8 separate populations that were each split into 4 technical replicate samples before collection and sequencing. Corruption was assessed as the rate of insertions of unexpected size. For 1 of the 8 populations, a correction to the corruption calculation was implemented, as aberrant long insertions created a duplicate copy of the recording locus, which then acquired subsequent, normal insertions (see Methods). With or without this correction, we saw very low rates of corruption (Figure 4.3e). At most 2% of records were corrupted per round. After 5 rounds of editing, 90% of loci remain uncorrupted (Figure 4.3e). Indeed, including all 8 populations, with no adjustments, only 0.34% of all sequences were corrupted.

In addition to faithfully accumulating edits, peCHYRON demonstrated that it acquires edits in an unbiased fashion. This deeply sequenced dataset above allowed us to calculate the information content of each edit: the observed Shannon entropy was 3.1 bits per A→B edit and 3.6 bits per B→A edit. This showed that pegRNAs with different signatures incorporated with little bias: for example, if 10 possible edits were incorporated with no bias, we would expect a Shannon entropy value of ~3.33 bits[25]. Thus, peCHYRON recording encoded information

supplied by pegRNAs with little bias, and retained the information efficiently, again suggesting that it is compatible with truly long-term recording.

4.3.4 Reconstruction of lineage relationships using peCHYRON

We applied peCHYRON to trace the lineage relationships among populations of cells descended from a parent population *via* a complex splitting process (Figure 4.4a). To achieve the maximum possible editing efficiency, we used serial transient transfections to drive high levels of prime editor and pegRNA expression. We transiently transfected one population of $293T_{6xHis}$ cells with plasmids encoding PE2 and 42 pegRNAs representing A→B and B→A pairs with varied signature mutations. After allowing the cells to grow for 8 days, we isolated 4 populations of 20,000 cells each. The next day we transfected again, allowed the cells to grow for 3 days, then split each population to yield 8 populations. Cells were transfected again the next day, then allowed to grow for 4 additional days, before being split again to yield 16 final populations. Two days later, ~1.6 million cells were collected from each population, DNA was extracted and the peCHYRON recording locus was subjected to amplicon sequencing. Three of the 16 populations were excluded from analysis so that the researchers performing the reconstruction could be blinded to both the identity of the wells and the exact shape of the lineage tree. From the sequences in the remaining 13 populations, we first filtered sequences by length (number of insertions), then by frequency of occurrence. We discarded all sequences with fewer than 4 rounds of insertion as the probability of generating a specific sequence of 3 signature mutations by chance rather than by descent was significant. Finally, we discarded sequences with frequencies of occurrence below a cutoff determined by the kneedle algorithm[26]. This represented a filter that removed spurious sequences that are likely library prep artifacts. After these two filtering steps, a total of 4571 sequences remained. For each pair of wells, we counted the number of shared identical peCHYRON sequences to calculate

Jaccard similarity[27] and then used the Jaccard similarity scores to perform agglomerative hierarchical clustering. The resulting tree accurately reconstructed all aspects of the splitting procedure (Figure 4.4b).

Calculating Jaccard similarity worked well for comparing populations of cells, but it only took into account identical peCHYRON sequences shared among the populations. Because the peCHYRON recording locus progressively acquires signature mutations in temporal order, related cells can share initial edits and then diverge. Sequences with shared early signatures but diverged late signatures can provide higher resolution information about lineage relationships, theoretically approaching single-cell-resolution. We sought to establish a lineage reconstruction algorithm that would take advantage of this ordered nature of recording. To do so, we modified the Jaccard similarity index to allow for partial matches. Pairs of populations were compared as follows. First, each sequence of signature mutations was split into a

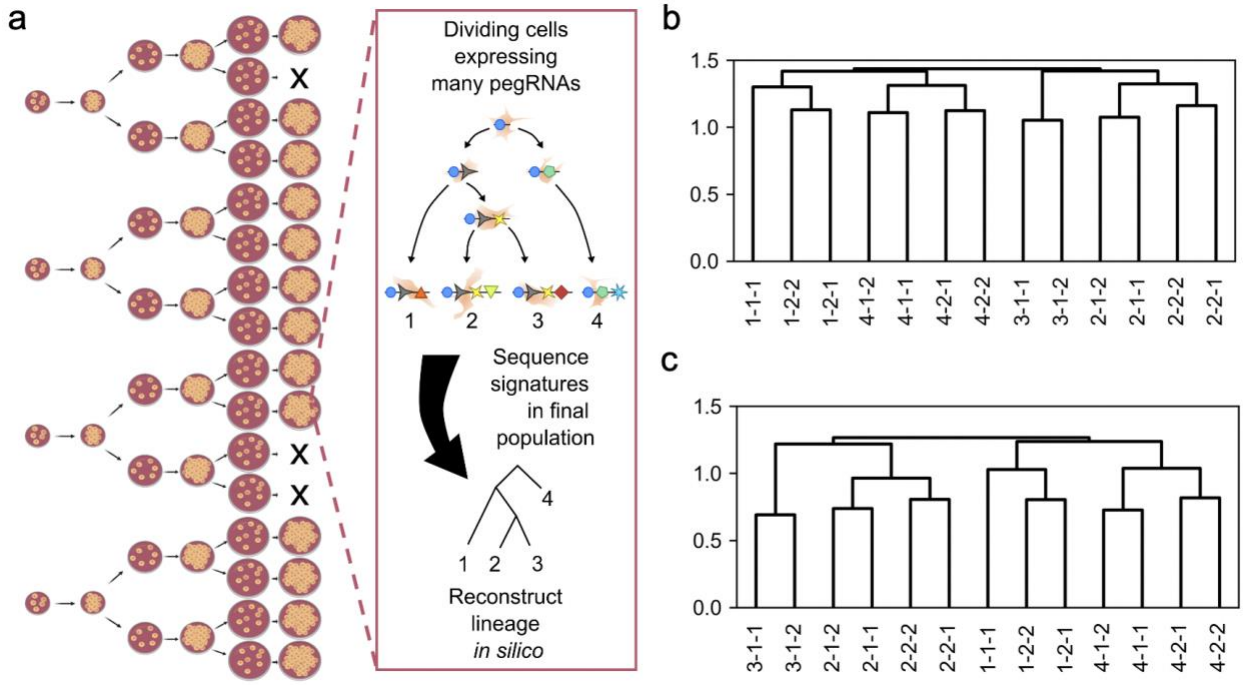


Figure 4.4. peCHYRON enables lineage tracing and temporally-resolved event recording. **a,** Culture splitting procedure for first lineage tracing experiment, using 293T cells transiently transfected with peCHYRON components. Inset shows how signature mutations accumulate in individual cells, then inform lineage reconstruction. **b,** Culture splitting patterns were accurately reconstructed using the Jaccard similarity index. **c,** Culture splitting patterns were accurately reconstructed using a modified Jaccard similarity index that accounts for sequences that share early signature mutations but diverge in later signature mutations.

collection of all possible subsequences that started from the first signature mutation (*i.e.,* prefixes). Each prefix generated from a sequence was given a fractional weight equal to the inverse of the number of prefixes making up the sequence. This resulted in a multiset of prefixes. We then computed a weighted multiset Jaccard similarity between all pairs of multisets. To illustrate this, consider a record ABCD where A, B, C, and D each represent a unique 3-bp signature mutation. This is split into 4 prefixes, A, AB, ABC, and ABCD, each given a weight of ¼. Now, consider a second record, ABCEF. This is split into 5 prefixes, A, AB, ABC, ABCE, and ABCEF, each given a weight of ⅕. When these are compared, the prefixes A, AB, and ABC match, and we add the minimum count, ⅕, for each match. The sum of these matches becomes the numerator for the Jaccard similarity, and the denominator is simply the sum of the cardinalities of the sets minus the numerator. With this new algorithm, the full splitting procedure was again accurately reconstructed (Figure 4.4c). Lineage reconstruction with this prefix Jaccard algorithm outperformed that done with traditional Jaccard similarity at almost every level of random downsampling and gave near-perfect reconstructions even when downsampling to only 20% of the data (Supplementary Figure 4.5), forecasting the utility of this reconstruction algorithm in realistic lineage tracing experiments where populations are poorly sampled.

4.3.5 Reconstruction of cellular event histories using peCHYRON

One of the major promises of ordered recording has been that biologists could trace the patterns of signaling that underlie the acquisition of specific cell fates. Ultimately, we would wish to read out the signaling history of single cells or at least subpopulations demarcated by simultaneous single-cell RNA sequencing. A more immediately achievable goal is to treat separate populations of cells as sensors of their environment over time by creating DNA recorders that make specific records in response to specific environmental signals. Ordered recording at CRISPR arrays in both bacterial populations[5,8,28] and single bacteria[29] has been

used to successfully report orders of signals experienced by the cells. To test a similar capability in mammalian cells, we used peCHYRON to record, then reconstruct, the order of transient events experienced by populations of 293T cells. Specifically, two types of transient events were recorded—pulsed expression of pegRNAs with distinct signature mutations, and multiplexed chemical induction of pegRNAs from inducible promoters. In both cases, histories could be reconstructed with high fidelity by reading out peCHYRON sequences (Figure 4.5 and Supplementary Figure 4.6).

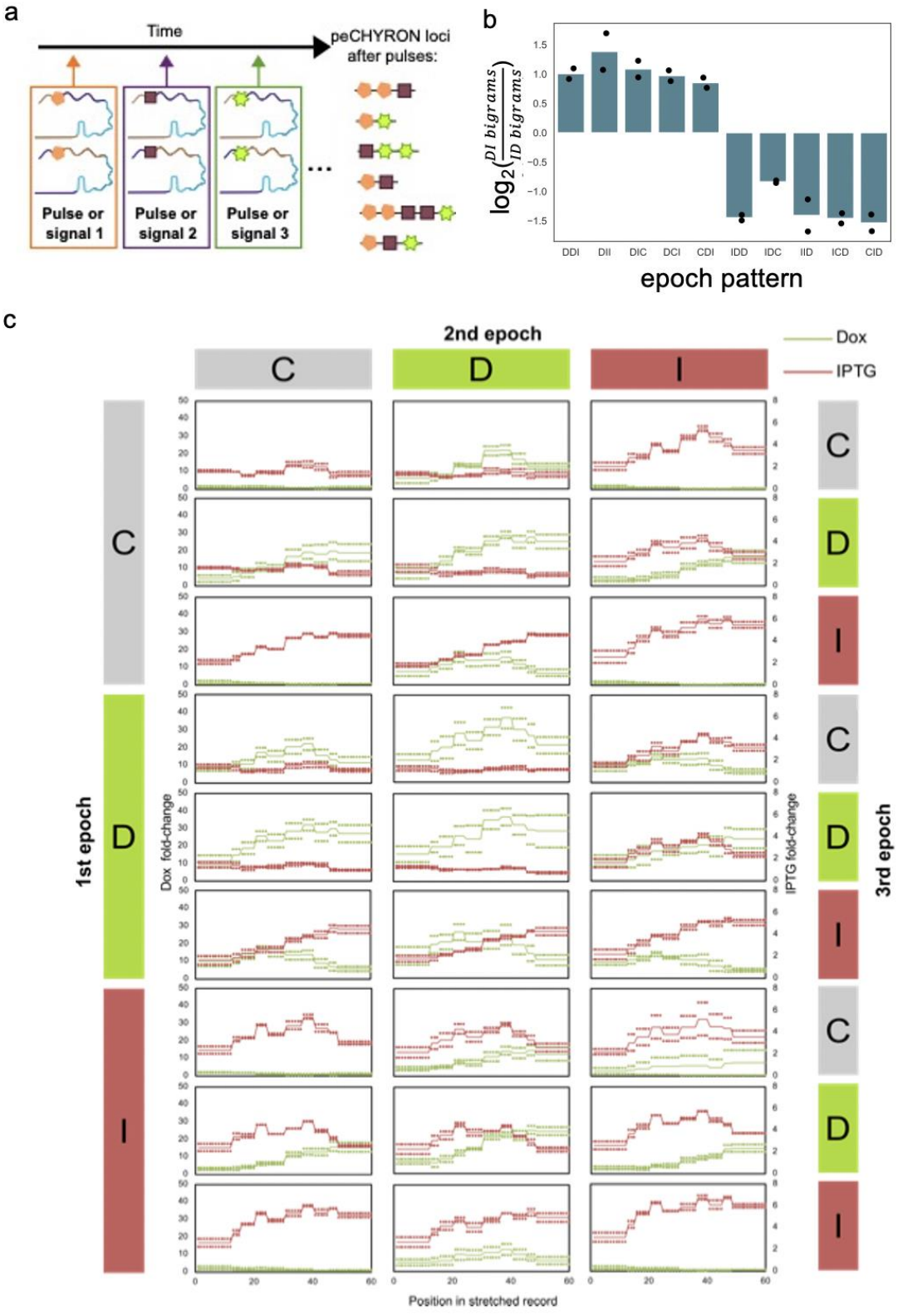

a
Time
peCHYRON loci after pulses:
Pulse or signal 1
Pulse or signal 2
Pulse or signal 3
b
$\log_2(\frac{DI\ bigrams}{ID\ bigrams})$
DDI DII DIC DCI CDI IDD IDC IID ICD CID
epoch pattern
c
2nd epoch
C
D
I
Dox
IPTG
1st epoch
3rd epoch
Dox fold-change
IPTG fold-change
Position in stretched record

Figure 4.5. peCHYRON enables temporally-resolved event recording. **a,** General approach to recording pulses (or any time-dependent signals) that are linked to expression of pegRNAs with known signature mutations. A pair of A→B and B→A pegRNAs are expressed together for any signal to be recorded, so that either pegRNA can record the signal, and both pegRNAs can continuously record the signal if it is present for a long duration. Colored shapes represent signature mutations. **b,** Ordered pairs of recorded signatures (bigrams) can be used to accurately determine the overall order of chemical exposures. $293T_{6xHis}$ cells in which PEmax, the protective nicking sgRNA, TetR, LacI, and pairs (A→B and B→A) of constitutive or inducible pegRNAs were stably integrated, then exposed to doxycycline (D), IPTG (I), or no inducer (constitutive only or C) for 3 epochs of 12 days each. After the 3rd epoch, recording loci were sequenced and the relative abundance of DI and ID signature mutation bigrams in the records were calculated. Points indicate biological replicate values. **c,** For the experiment also shown in (b), The signature mutations were extracted from peCHYRON records, then the sequence of signature mutations from each recording locus were computationally stretched to a constant length (this is 60 because it is the least common multiple of all insertion lengths in the data set). Stretched records were pooled together in a manner that allows each insertion length to contribute equally to the final pool (*i.e.*, long insertions are weighted more heavily to prevent being overshadowed by short, abundant insertions that are less informationally rich). At every position along the stretched records, the proportions of each signature were calculated. The fold change of this proportion relative to a control in which cells were exposed to no inducer for 18 days was plotted for doxycycline (D, green) and IPTG (I, red). Points represent biological replicates and lines are the mean of the biological replicates. Data is organized according to the inducer, or lack thereof, that each sample was exposed to during the 3 epochs.

To interpret peCHYRON records, a previously reported bigram method[15] and a novel "signature stretching" method were employed (Figure 4.5a and Supplementary Figure 4.6a). For simplicity, consider two events, X and Y, each of which promotes expression of a set of pegRNAs that install a signature X or Y. The order of signature mutations recorded in peCHYRON loci can reveal the order in which X and Y occurred, since each signature mutation reflects the identity of an exact pegRNA, and the sequential nature of edit accumulation preserves temporal information. In many cases, the bigram method is appropriate for interpreting the order of events. For this, the frequency of ordered pairs of signatures is calculated across all recording loci in a population of cells. The frequency with which X appears before Y, or vice versa, in individual records, should correspond to the order of the events: if X occurred before Y, we would expect the bigram XY to be more abundant than YX (Figure 4.5a and Supplementary Figure 4.6a). Although bigram frequencies have been previously reported to successfully recapitulate the order of pegRNA pulses[15], an inherent limitation of this approach is that it cannot resolve repeated pulses of a given signature. For example, if a population of cells experienced event X, then Y, then X again at a later time, the bigram frequencies would show high incidence of both XY and YX bigrams, with no way to discern how many times the cells

alternated between expression of X and Y (*i.e.*, events ordered X-Y-X would be indistinguishable from X-Y-X-Y, X-Y-X-Y-X-, etc.). To overcome this, we developed a new analysis approach that leverages the fundamental relationship between timing of events and position in the peCHYRON recording locus. All records with 2 or more insertions contain some information about the true order of events, so we aggregate this information by computationally "stretching" the signatures from all records with 2+ insertions to the same length, then calculating the fraction of each signature at each position among the pool of stretched records. We then calculate the enrichment of each signature relative to controls with no inducers present. For example, if all records are stretched to a length of 60, a record with two insertions with signatures XY would be stretched to 30 Xs followed by 30 Ys. A record with three insertions XYX would be stretched to 20 Xs, then 20 Ys, then 20 Xs (Supplementary Figure 4.6a). Longer insertions are intrinsically more informative because they originate from cells that recorded events more frequently throughout the course of the experiment; so, when calculating the fraction of signatures at each position, records are weighted proportionally to their length, such that records of each length contribute equally to the final pool of stretched records. Thus, we have two approaches that can be used to reconstruct the order of cellular events: bigrams, which take into account the relative positions of signatures on the same recording locus, and stretched records, which use population averages but are compatible with repeating event patterns.

We first validated peCHYRON's ability to perform temporally resolved event recording by reconstructing pulsed expression of pegRNAs in a population of cells. For this, we used $293T_{6xHis}$ cells into which PEmax and the protective nicking guide had been stably integrated. We split this cell line into 2 separate populations, then used a random number generator to choose a pair of A→B and B→A pegRNAs to transfect into each population. Both members in the pair of pegRNAs install the same signature mutation as each other. With this setup, a pair of pegRNAs is included to enable propagation with alternating A→B and B→A steps, although

cells that acquire only a single insertion successfully record the presence of the pegRNA pulse. We repeatedly chose random pegRNA pairs to transfect, for a total of 4 rounds of sequential transfections. Cells were transfected every six days or every three days. Once used, a pegRNA pair was never repeated. We collected cells at the end of the experiment, extracted DNA, and subjected the peCHYRON locus to amplicon sequencing. As expected, because each signature mutation was only used once, the order of transfections in our experiment can be inferred correctly by calculating the frequencies of all bigrams (Supplementary Figure 4.6b). When the records were computationally stretched so that proportions of each signature at each position could be calculated, the true order of transfections was again reconstructed (Supplementary Figure 4.6c).

We also created cells that could sense and record patterns of chemical signals, without a need for interventions such as transfection. Again, the bigram and “stretching” analysis methods were used to demonstrate successful recording of events. After optimizing the parts, we made cell lines that express PEmax and a protective nicking guide, as well as three different pairs of A→B and B→A pegRNAs with different, known signatures: one pair constitutively, one pair in response to doxycycline, and one pair in response to IPTG. All necessary constructs were transfected as PiggyBac cargo, along with a transient PiggyBac transposase plasmid. Integrants were selected with puromycin (which marks PEmax) and blasticidin (which marks the protective nicking sgRNA). Six days after the transfection, cells were passaged into medium containing doxycycline, IPTG, or no drug. This first “epoch” lasted for 12 days, during which cells were passaged as needed. We completed a total of 3 epochs, each lasting 12 days. We tested 27 exposure patterns, representing all possible combinations of doxycycline exposure, IPTG exposure, and no treatment, with two biological replicates of each exposure pattern. Records were read out by amplicon sequencing of DNA from 50,000-150,000 cells per sample. For the 10 exposure patterns in which doxycycline was added either before or after IPTG, but not both, the relative frequencies of dox-IPTG and IPTG-dox bigrams corresponded to the true

order of exposure (Figure 4.5b), marking peCHYRON's success in distinguishing between the order of two induced signals. However, the comprehensive set of 27 exposure patterns included repeated instances of each inducer (or lack thereof), which is the use case that cannot be resolved with bigram frequencies and instead benefits from our analysis method that computationally stretches the records. When the records were computationally stretched, the proportion of edits at each position, compared to a no-inducer control, again successfully reflected the order of the doxycycline and IPTG treatments for all samples (Figure 4.5c). In addition to the relative order of induction events, the absolute timing of induction was also captured in the records. For example, samples that were exposed to doxycycline in the first or second epoch showed enrichment for the corresponding signature in the early or middle positions of the stretched records, whereas samples that were induced in the third epoch exhibited a later increase in the associated signature. The same was true for IPTG induction. In addition, the relative durations of induction applied to each sample were often distinguishable, allowing for comparisons of cells' experiences among samples in the same experiment. Although the editing outcomes for each pattern of induction followed expected trends, we acknowledge that the differences between stretched records for similar epoch patterns were nearly identical in some cases (*e.g.*, two epochs of IPTG and three epochs of IPTG). In the future, higher-efficiency editing by peCHYRON will enhance its ability to distinguish similar but not identical patterns of cell history.

For use of peCHYRON to record transient cellular events *in vivo*, the most obvious method for interpreting records would be to perform experiments *in vitro* with known durations and known orders of the events, then compare those *in vitro* records to the *in vivo* records. To make it possible to determine to what extent the current version of peCHYRON could be used in this way, we compared the stretched records from the two biological replicates for the experiment shown in Figure 4.5 to each other. As a better mimic of replicates in slightly different cell types, the amount of PEmax plasmid added when creating stable cell lines for the replicates

was varied by 10%. We computed the pairwise Manhattan distances between the compositions of stretched records in each sample in replicate one and each sample in replicate two (see Methods). Ideally, identical epoch patterns would have a very low distance, and distances would increase as fewer epochs are shared between samples. There was a significant correlation between Manhattan distances and shared epochs between samples, with consistently low Manhattan distances for replicates that shared the same overall pattern of epochs. However, there was overlap in the distributions of distances between identical and unrelated epoch patterns, meaning that some epoch patterns would be incorrectly assigned based on the replicate experiment (Supplementary Figure 4.7). Thus, peCHYRON records are correlated with the timings of two types of chemical exposures relative to each other, and to no treatment, including in situations in which exposures alternate, but the records do not yet capture events at high enough resolution to draw quantitative conclusions about cells' experiences. Still, the current version of peCHRYON allows qualitative interpretation of the stretched records, which can shed light on the relative experiences of populations of cells.

4.4 Discussion

Our results establish peCHYRON as an advanced DNA recorder that autonomously generates ordered, high-information records *in vivo*. These records are exceptionally durable, can be parsed to decode the temporal order of recorded events, and can be arbitrarily long since the rate of sequential insertions at the peCHYRON locus does not slow as the locus increases in length. In keeping with these favorable performance characteristics, peCHYRON was used to accurately reconstruct complex cell lineage relationships and event histories in proof-of-concept experiments that exploit the information available in temporally ordered records.

For a DNA recorder to continuously operate in the context of a developing organism, it needs to perform robustly in a variety of cell types, as the cells that are being tracked undergo phenotypic changes. One major source of variation in genome editing outcomes between cell

types arises from each host cell's complement of DNA repair proteins[30]. Prime editing is enabled by widely-expressed DNA replication factors[23,31,32] and can be antagonized by mismatch repair (MMR)[23] and double-strand break (DSB) repair. DSB repair (*e.g.*, non-homologous end joining and homologous recombination) can lead to unintended indels when both DNA strands are nicked[33], as in some prime editing strategies (*i.e.,* PE3[21]). However, because only one strand is nicked during prime editing with the PE2 architecture used by peCHYRON, we do not expect the DSB repair pathway to be triggered during recording. In agreement with this model, very low rates of unintended indels are created with PE2 ([21] and Figure 2a). The other known pathway that typically curbs prime editing, mismatch repair (MMR), is deficient in 293T cells[34]. Thus, prime edits that are highly efficient in 293Ts often perform substantially worse in other cell types[23]. To peCHYRON's benefit, however, insertions longer than 13 bp are not subject to MMR, even in MMR-competent cells[35], meaning our 20-bp insertions are unlikely to be affected. The efficiency of smaller prime edits that are subject to MMR has been increased by expressing a dominant negative mutant of *MLH1*[23], but implementing this MLH1dn strategy in an animal model would likely interfere with the biology we seek to study. Due to its minimal interface with host DNA repair, peCHYRON is exceptionally well-positioned for consistent activity across cell types, without disrupting the native DNA repair milieu in the cells of interest.

It is instructive to compare peCHYRON to DNA Typewriter[15], which was pre-printed simultaneously with peCHYRON and has subsequently been published. DNA Typewriter uses prime editor to install 6-bp insertions in temporal order on pre-defined recording arrays. Key advantages of DNA Typewriter over peCHYRON include relative simplicity of optimization and part expression (*i.e.*, only one type of pegRNA and no protective nicking guide), and high efficiency (*e.g.*, in a lineage tracing experiment in a 293T clonal cell line, the majority of loci recorded at least four edits)[15]. Key advantages of peCHYRON over DNA Typewriter include evasion of MMR, and the ability to add its own target sites, opening up the possibility of editing

at many sites in parallel in highly repetitive mammalian genomes[9] and allowing in principle arbitrarily many rounds of recording. In addition, peCHYRON is likely to have an advantage in contending with the inherent instability of highly repetitive DNA arrays[15,36], since repetitive units are added only as records are made in peCHYRON rather than being present throughout the experiment as in DNA Typewriter. Also, the periodicity of repeating sequences in peCHYRON is 40 bp, compared to 14-20 bp for DNA Typewriter, so that acquisition of an equal number of edits with peCHYRON and DNA Typewriter would result in ≥ 2-fold more repeated sequence motifs in DNA Typewriter than in peCHYRON. The peCHYRON architecture represents a useful addition to the DNA recording toolkit, especially for applications that call for indefinite expansion of the recording locus (*e.g.*, recording over very long timespans).

In the near term, we aim to improve the resolution of recorded lineage phylogenies and the confidence of event history reconstructions by increasing the efficiency of editing with peCHYRON. We may achieve this *via* optimizing the propagator sequences, especially of the B→A pegRNA. However, even with the current editing efficiencies, the central architecture of peCHYRON possesses the critical features of a DNA recorder well-suited to long-term experiments. For example, deep lineage tracing on the scale of a whole mammal requires a recorder that can continuously operate throughout development and generate durable records capable of distinguishing among billions of cells[37]. peCHYRON can already continuously operate for at least 2 months without decreases in the rates of propagation, and it encodes 3.1-3.6 bits of information in each sequentially inserted record such that a record containing just 11 signature mutations has ~40 billion possible states, similar to the number of cells in an adult mouse. peCHYRON is exceptionally well-matched for temporally-resolved multiplexed signal recording, as it accumulates records in an ordered manner, and the number of signals it can record can be expanded by expressing pegRNAs with known signature mutations in response to promoters of interest. Here, we have recorded the orders of chemicals added to cell

populations and have demonstrated our pegRNAs can tolerate at least 30 unique 3-nt signature mutation sequences, each of which could be theoretically linked to an additional inducible promoter to record highly complex histories of what cells experience *in situ*. Applications of peCHYRON beyond the proof-of-concept experiments shown here will capitalize on these and other opportunities in the quest to understand the detailed histories of individual cells in animal biology and development.

4.5 References

1. Sheth, R. U. & Wang, H. H. DNA-based memory devices for recording cellular events. *Nat Rev Genet* 19, 718–732 (2018).

2. McKenna, A., Findlay, G. M., Gagnon, J. A., Horwitz, M. S., Schier, A. F. & Shendure, J. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Sci New York N Y* 353, aaf7907 (2016).

3. Kalhor, R., Kalhor, K., Mejia, L., Leeper, K., Graveline, A., Mali, P. & Church, G. M. Developmental barcoding of whole mouse via homing CRISPR. *Sci New York N Y* 361, eaat9804 (2018).

4. Schmidt, S. T., Zimmerman, S. M., Wang, J., Kim, S. K. & Quake, S. R. Quantitative analysis of synthetic cell lineage tracing using nuclease barcoding. *Acs Synth Biol* (2017). doi:10.1021/acssynbio.6b00309

5. Sheth, R. U., Yim, S. S., Wu, F. L. & Wang, H. H. Multiplex recording of cellular events over time on CRISPR biological tape. *Sci New York N Y* 358, 1457–1461 (2017).

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

7. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Sci New York N Y* 353, aaf1175 (2016).

8. Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* 562, 380–385 (2018).

9. Hwang, B., Lee, W., Yum, S.-Y., Jeon, Y., Cho, N., Jang, G. & Bang, D. Lineage tracing using a Cas9-deaminase barcoding system targeting endogenous L1 elements. *Nat Commun* 10, 1234 (2019).

10. Chan, M. M., Smith, Z. D., Grosswendt, S., Kretzmer, H., Norman, T. M., Adamson, B., Jost, M., Quinn, J. J., Yang, D., Jones, M. G., Khodaverdian, A., Yosef, N., Meissner, A. & Weissman, J. S. Molecular recording of mammalian embryogenesis. *Nature* 570, 77–82 (2019).

11. Bowling, S., Sritharan, D., Osorio, F. G., Nguyen, M., Cheung, P., Rodriguez-Fraticelli, A., Patel, S., Yuan, W.-C., Fujiwara, Y., Li, B. E., Orkin, S. H., Hormoz, S. & Camargo, F. D. An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells. *Cell* (2020). doi:10.1016/j.cell.2020.04.048

12. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & Oudenaarden, A. van. Whole-organism clone tracing using single-cell sequencing. *Nature* 556, 108–112 (2018).

13. Chow, K.-H. K., Budde, M. W., Granados, A. A., Cabrera, M., Yoon, S., Cho, S., Huang, T., Koulena, N., Frieda, K. L., Cai, L., Lois, C. & Elowitz, M. B. Imaging cell lineage with a synthetic digital recording system. *Science* 372, eabb3099 (2021).

14. Loveless, T. B., Grotts, J. H., Schechter, M. W., Forouzmand, E., Carlson, C. K., Agahi, B. S., Liang, G., Ficht, M., Liu, B., Xie, X. & Liu, C. C. Lineage tracing and analog recording in mammalian cells by single-site DNA writing. *Nat Chem Biol* 17, 739–747 (2021).

15. Choi, J., Chen, W., Minkina, A., Chardon, F. M., Suiter, C. C., Regalado, S. G., Domcke, S., Hamazaki, N., Lee, C., Martin, B., Daza, R. M. & Shendure, J. A time-resolved, multi-symbol molecular recorder via sequential genome editing. *Nature* 608, 98–107 (2021).

16. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science* 353, aag0511–aag0511 (2016).

17. Kalhor, R., Mali, P. & Church, G. M. Rapidly evolving homing CRISPR barcodes. *Nat Methods* 14, 195–200 (2017).

18. Frieda, K. L., Linton, J. M., Hormoz, S., Choi, J., Chow, K.-H. K., Singer, Z. S., Budde, M. W., Elowitz, M. B. & Cai, L. Synthetic recording and in situ readout of lineage information in single cells. *Nature* 541, 107–111 (2017).

19. Tang, W. & Liu, D. R. Rewritable multi-event analog recording in bacterial and mammalian cells. *Sci New York N Y* 360, eaap8992 (2018).

20. Chen, W., Choi, J., Nathans, J. F., Agarwal, V., Martin, B., Nichols, E., Leith, A., Lee, C. & Shendure, J. Multiplex genomic recording of enhancer and signal transduction activity in mammalian cells. *Biorxiv* 2021.11.05.467434 (2021). doi:10.1101/2021.11.05.467434

21. Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., Chen, P. J., Wilson, C., Newby, G. A., Raguram, A. & Liu, D. R. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576, 149–157 (2019).

22. Loveless, T. B., Carlson, C. K., Hu, V. J., Helmy, C. A. D., Liang, G., Ficht, M., Singhai, A. & Liu, C. C. Molecular recording of sequential cellular events into DNA. doi:10.1101/2021.11.05.467507

23. Chen, P. J., Hussmann, J. A., Yan, J., Knipping, F., Ravisankar, P., Chen, P.-F., Chen, C., Nelson, J. W., Newby, G. A., Sahin, M., Osborn, M. J., Weissman, J. S., Adamson, B. & Liu, D. R. Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. *Cell* 184, 5635-5652.e29 (2021).

24. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533, 420–424 (2016).

25. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst Tech J* 27, 379–423 (1948).

26. Koeppel, J., Weller, J., Peets, E. M., Pallaseni, A., Kuzmin, I., Raudvere, U., Peterson, H., Liberante, F. G. & Parts, L. Prediction of prime editing insertion efficiencies using sequence features and DNA repair determinants. *Nat Biotechnol* 1–11 (2023). doi:10.1038/s41587-023-01678-y

27. Mathis, N., Allam, A., Kissling, L., Marquart, K. F., Schmidheini, L., Solari, C., Balázs, Z., Krauthammer, M. & Schwank, G. Predicting prime editing efficiency and product purity by deep learning. *Nat Biotechnol* 1–9 (2022). doi:10.1038/s41587-022-01613-7

28. Satopää, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a “Kneedle” in a Haystack: Detecting Knee Points in System Behavior. *2011 31st Int Conf Distributed Comput Syst Work* 166–171 (2011). doi:10.1109/icdcsw.2011.20

29. Jaccard, P. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. *New Phytol* 11, 37–50 (1912).

30. Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P. & Canaider, S. An estimation of the number of cells in the human body. *Ann Hum Biol* 40, 463–471 (2013).

CHAPTER 5: Development of inducible prime editing guide RNAs toward multiplexed signal recording with peCHYRON

5.1 Introduction

Every pegRNA in the peCHYRON architecture produces a predictable edit at the recording locus. The well-defined nature of the edits allows direct linkage between user-defined inducers and known signature mutations in the pegRNAs, such that the presence of a given signature mutation in the recording locus can be used to deduce the presence of a known inducer. With peCHYRON's ordered nature of recording and clear punctuation between sequential edits, this architecture could be extremely powerful for logging cells' experiences over time. A theoretical maximum of 64 signals could be recorded simultaneously (each signal would be encoded by a propagating pair of pegRNAs with unique 3-bp signature mutations— both pegRNAs in the propagating pair for a given signal would have the same signatures as each other, but all propagating pairs would be different from each other), and there is no known upper limit on the duration of recording with peCHYRON. The punctuation between edits enables precise measurement of how many edits have occurred, which reduces ambiguity when interpreting analog records.

Achieving analog signal recording with peCHYRON is challenging because in nature, eukaryotic cells use RNA polymerase III (RNAP III) to transcribe untranslated RNAs (*e.g.*, tRNAs, rRNAs). RNAP III is well-suited for production of these small, highly structured RNAs because it initiates and terminates transcription at very precise positions, and it doesn't modify the resultant RNA with a 5' cap or polyA tail[1]. Thus, when small and structured RNAs are required in engineered eukaryotic cells, promoters that recruit RNAP III for transcription are typically used. Cas9 guide RNAs (sgRNAs) and prime editing guide RNAs (pegRNAs) are firmly in this category of engineered RNAs that are typically transcribed by RNA polymerase III (RNAP

III). However, all natural pol III promoters are constitutive, which renders this mode of expression inappropriate for signal recording. A few synthetic inducible RNAP III promoters exist[2], but there is a large and ever-growing toolbox of inducible promoters for RNA polymerase II (RNAP II)[3,4]. Also, we want users of peCHYRON to be empowered to record the activity of endogenous inducible promoters, all of which recruit RNAP II. Thus, for peCHYRON to serve as a versatile measuring device for biologically interesting signals, the pegRNAs must be expressed with RNAP II. Herein lies the challenge, since the native role of RNAP II is transcribing mRNA, which necessarily has a 5' cap and poly(A) tail[5]. The 5' cap interferes with Cas9 editing by sterically hindering target recognition, plus the presence of these motifs causes transcripts to be exported to the cytoplasm[1]. The recording locus for peCHYRON is in the genome, so of course, trafficking the pegRNAs out of the nucleus is counterproductive.

Before the advent of peCHYRON, Nissim et al. developed several RNA processing schemes to cleave sgRNAs out of RNAP II-transcribed mRNAs [1]. We used design principles from their work to guide our own engineering efforts for RNAP II-expressed pegRNAs (Figure 5.1a). This chapter describes the steps that have been taken to establish RNAP II-expressed pegRNAs for inducible signal recording with peCHYRON. Further development is needed to tie up the loose ends this story, but I intend for this information to aid in the next engineering steps.

5.2 Methods

5.2.1 Molecular cloning

A plasmid that encodes constitutive expression of prime editor (PE2*)[6] was a gift from Eric Wolfe, and it was used directly for transient transfections in HEK293Ts, or PCR-amplified and inserted into a PiggyBAC cargo vector for stable cell line creation. A constitutive Cas9 expression plasmid was a gift from George Church (Addgene # 41815), and it too was used directly for genome editing experiments. To achieve hypoxia-dependent gene expression, the

4xHRE_YB_TATA promoter was PCR amplified from 4xHRE_v2_YB-TATA-Gluc-CMV_dsRed[7] (Addgene #117399), which was a gift from Yvonne Chen. The 7xTCF-LEF promoter for Wnt inducible gene expression was PCR-amplified from the M50 Super 8x TOPFlash reporter plasmid[8] (Addgene # 12456), gifted by Randall Moon. For doxycycline-inducible gene expression, the TetON-3G promoter (Addgene # 96963) was used, with a V16 mutation for experiments involving stable cell lines and V10 mutation for experiments involving transient transfections[9].

Constitutive pegRNA expression plasmids were constructed by modifying the sequences of a pcDNA3.1 plasmid[10] (Addgene # 126450) and a pGL2-Luc pegRNA plasmid[1] (Addgene # 55201), the latter of which was gifted from Timothy Liu. These vectors were used for transient transfections. For creation of stable cell lines, pegRNAs were cloned into a Part 0 plasmid from the Mammalian Toolkit[11] for a piggyBAC expression vector (Addgene # 1000000180). This was a gift from Hana El-Samad. For each pegRNA variant, custom sequence motifs were added to these plasmids *via* standard Gibson Assembly or restriction digestion and ligation (digestion with HF- enzymes, ligation with T4 DNA Ligase, both from NEB). After confirming faithful assembly *via* Sanger sequencing (Genewiz) or whole-plasmid sequencing (Primordium Labs), the pegRNA plasmids were hp-miniprepped (GenElute HP Plasmid Miniprep Kit, NA0150) to produce high-purity plasmid stocks for downstream transfections in HEK293T cells.

5.2.2 Cell culture

HEK293Ts were acquired from ATCC (CRL-3216) but not otherwise authenticated. In all cases, cell culture was conducted in DMEM with high glucose plus GlutaMAX (Gibco, 10566024) and 10% FBS (Sigma, 12306C). Cultures were maintained at 37 °C and with 5% CO2.

To assay individual pegRNA expression designs, HEK293Ts were thawed from our liquid nitrogen collection, using a bead bath at 37 °C for approximately 5 minutes. Cells were gently resuspended in 37 °C DMEM + FBS and plated into cell culture treated dishes. 1-2 days

later, cultures were split into 6-well dishes (Falcon, 353046) at a 5% seeding density for transient transfections, or 24-well dishes at a 20% seeding density for establishing stable cell lines with PiggyBAC transposase.

For transient transfections, hp-minipreps of prime editor (1500 ng) and a pegRNA (500 ng), or Cas9 (1500 ng) and a sgRNA (500 ng) were transiently transfected using FuGENE HD Transfection Reagent (Promega E2312), following manufacturer instructions. 20-minute complexing times with FuGENE were used. 3 days post-transfection, the cells were collected by aspirating the media out of the cultures, spraying cells off the plates with cold PBS, pelleting the cells, removing the supernatant, and placing cell pellets at -80°C for later processing.

For experiments involving stable cell lines, hp-minipreps of the piggyBAC cargo plasmids were co-transfected with Super PiggyBAC Transposase (SBI, System Biosciences) with FuGENE HD Transfection Reagent (Promega E2312), following manufacturer instructions. 20-minute complexing times with FuGENE were used. One day after transfecting, the cultures were passaged into a larger culture dish (6-wells). Approximately 24 hours later, the media was replaced with antibiotic-containing DMEM + 10% FBS (either 10 ug/mL blasticidin or 2 ug/mL puromycin) to select for stable integrants. Puromycin selection was conducted for 2 days, at which point untransfected control cells completely died off. Blasticidin selection was conducted for 10 days. In cases where stable cell lines were later transiently transfected with additional constructs, the antibiotic selection step was completed, and the antibiotic was removed from the media prior to transfecting as described above.

To induce doxycycline-, hypoxia-, and Wnt-responsive pegRNA constructs, doxycycline was added to the culture media at a final concentration of 1 ug/mL. Hypoxia was simulated with the chemical hypoxia mimic, DMOG, which was added to the media at a final concentration of 1 mM. Similarly, Wnt signaling was mimicked with a chemical effector LiCl at a final concentration of 20 mM. For transiently transfected cultures, the transfection media was aspirated off the cells and replaced with induction media 12-16 hr post-transfection.

5.2.3 Sample preparation for next-generation sequencing

Frozen cell pellets were subjected to genomic DNA (gDNA) extraction with a QIAamp DNA Mini kit (Qiagen, 51304). After DNA extraction, the appropriate region was amplified by PCR. The primers contained partial Illumina adapters and a 5 – 7 nt sample-specific barcode for demultiplexing. The PCR was performed with Phusion HotStart Flex DNA Polymerase with GC buffer, with or without DMSO (NEB), 30-35 cycles. Before sequencing on an Illumina flow cell, reactions were pooled and purified by binding to AMPure beads (0.9 beads:1 sample). The libraries were sent to Genewiz, Inc. for Amplicon-EZ sequencing, where they were further amplified to incorporate the TruSeq HT i5 and i7 adaptors and then sequenced on an Illumina HiSeq 2500 with a paired-end 250 protocol.

5.2.4 Computational analysis of DNA sequences

The sequences retrieved by HTS were first demultiplexed based on their barcodes. The script is available at github.com/liusynevolab/CHYRON-NGS. After demultiplexing, the forward and reverse fastq files for each sample were inputted to the CRISPResso2 web tool, which quantified the abundance of each insertion length. The overall insertion lengths were used to calculate editing efficiency.

5.3 Results

5.3.1 Initial attempts using self-cleaving ribozymes to liberate pegRNAs from RNAPII transcripts yielded little to no editing

We set out to achieve pegRNA expression from mRNAs with an approach that wouldn’t require expression of additional enzymes. One of the RNA processing techniques demonstrated by Nissim et al. was the addition of self-cleaving ribozyme sequences on either side of the guide

RNA (hammerhead ribozyme on the 5' end and HDV ribozyme on the 3' end of the sgRNA)[1]. With this, the ribozyme sequences alone can catalyze removal of the sgRNA from the mRNA with surgical precision. We tested ribozyme-flanked pegRNAs, using a constitutive promoter to determine if the RNA processing approach was efficient. Although the purpose of expressing pegRNAs with RNAP II is to establish inducible pegRNAs, we reasoned that a well-characterized constitutive promoter with high expression would provide the best dynamic range for prime editing rates (as opposed to a less well-behaved inducible promoter). In other words, with a high expression promoter, if the self-cleaving ribozymes were functional, but moderately inefficient at liberating pegRNAs from the transcripts, we would still be likely to detect editing. This would be more informative than if we used an inducible promoter than had low expression.

To start, we straightforwardly substituted 6xHis-inserting pegRNA genes, flanked by self-cleaving ribozyme sequences, into a constitutive RNAP II expression vector that we had successfully used in many past experiments for protein expression (pcDNA3.1 with a CMV promoter[10]; we have firsthand success using it to express TdT, GFP, and other proteins in HEK293Ts). The pegRNA inserts an 18-nt 6xHis sequence because, at the start of this project, we had characterized this sequence as a likely component for peCHYRON propagation, but we had not yet settled on final peCHYRON pegRNA designs that make 20 nt insertions. The pcDNA3.1 vector has an SV40 origin of replication, which promotes episomal plasmid replication *via* the large T antigen in HEK293Ts[12]. We hypothesized that this would help achieve high expression of our pegRNAs, since plasmid copy number would be greater than that of a non-replicating vector. However, we noted that Nissim et al., who had previously achieved expression of ribozyme-flanked sgRNAs from RNAP II transcripts [1], used a vector without an SV40 origin for their experiments (pGL2-Luc). Inspecting the sequence motifs in the pGL2-Luc vector (Addgene #55200) didn't reveal any obvious advantages inherent to this plasmid, but we tested it in parallel with the pcDNA3.1 designs since it had been successfully used for sgRNA expression by others. As a positive control, we included an sgRNA gene in the pGL2-Luc

vector, and we also tested the same sgRNA gene in the pcDNA3.1 vector. In addition, we tested two different versions of the hammerhead ribozyme on the 5' end of the pegRNAs and sgRNAs—HHLu, which is an exact copy of the sequence used by Nissim et al., and HH1, which is a sequence that we designed, ourselves (the hammerhead ribozyme sequence is largely degenerate with the overall structure being the important factor). All constructs used an identical CMV enhancer and promoter pair, and all contained a gene for a fluorescent protein upstream of the guide RNA. The pGL2-Luc vector additionally harbored a triplex RNA sequence (a stabilizing tertiary structure that allows the fluorescent protein to be translated despite its poly(A) tail being cleaved off). We did not include the triplex in our pcDNA3.1 cassette because fluorescent protein readout wasn't prioritized for this experiment. The gene for the fluorescent protein was only included in case it affected the stability of the mRNA prior to cleaving out the pegRNA. As shown in Figure 5.1b, the sgRNAs yielded robust editing in all sequence contexts (63-82% as efficient as the RNAP III-expressed sgRNA control). The pGL2-Luc and pcDNA3.1 vectors produced similar overall levels of editing. On the other hand, the pegRNAs yielded much lower levels of editing in the best case (21% as efficient as the RNAP III-expressed pegRNA control). Both hammerhead ribozyme variants, HHLu and HH1, performed similarly. The expression vector, however, appeared to have a significant impact on expression. No editing was detectable with the pcDNA3.1 vector. Although this experiment was conducted in a single replicate, the results are supported by previous attempts to express other pegRNAs (with insertion sequences unrelated to peCHYRON propagation, but known to be efficient when expressed with RNAP III) also yielded no detectable editing (data not shown). Perhaps the triplex RNA sequence affects folding of the 5' hammerhead ribozyme. Therefore, we continued with the pGL2-Luc vector for subsequent experiments.

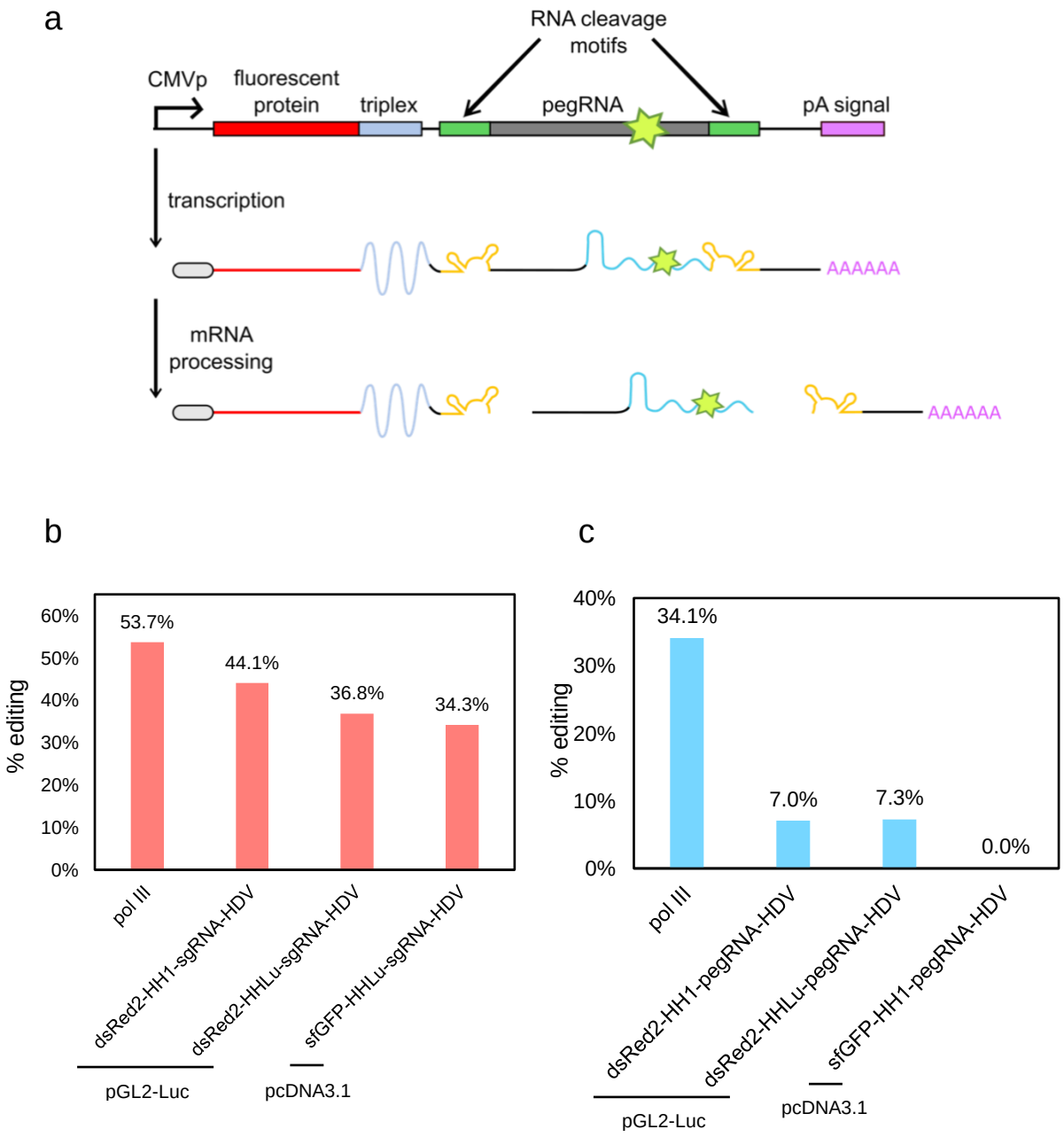


Figure 5.1. Initial attempts to express guide RNAs from RNAP II transcripts. **a,** Schematic of the RNA-processing technique for pegRNA expression. To settle on a functional technique for cleaving pegRNAs out of the transcripts, a constitutive CMV promoter was used (CMVp). The transcript contained a fluorescent protein gene, a triplex RNA sequence to facilitate expression of fluorescent protein, and the pegRNA flanked by sequences that lead to RNA cleavage. The pegRNA is precisely cleaved out of the mRNA. **b,** Initial attempts to express a sgRNA via RNAP II. The sgRNA was flanked by self-cleaving ribozymes sequences. HH1, hammerhead ribozyme sequence 1. HHLu, a hammerhead ribozyme sequence previously used by Nissim et al. HDV, hepatitis delta virus ribozyme. pGL2-Luc and pcDNA3.1 are two different mammalian expression vectors. **c,** A pegRNA that inserts a 6xHis tag was flanked by self-cleaving ribozymes and transcribed by RNAP II.

5.3.2 Appending structured sequence motifs to the 3' end of the 6xHis-inserting pegRNA improved editing efficiency for ribozyme-flanked constructs

The differences in editing efficiencies between the ribozyme-flanked sgRNAs and pegRNAs were surprising. These two types of guide RNAs are identical to each other, except the 3' end of the sgRNA is comprised of hairpins that form the scaffold for binding with Cas9, whereas the pegRNA has a 3' extension added to the scaffold[13]. Therefore, since the 3' end is the only distinction, we suspected the 3' extension on the pegRNA could have interfered with ribozyme folding. To address this, we attached the hairpin sequences that reside on the 3' end of the sgRNA onto the 3' end of the pegRNA, so the structure of the pegRNA's 3' end would hopefully mimic that of the sgRNA's 3' end. The overall sequence of the pegRNA in this design is [spacer]-[hairpins for scaffold]-[3' extension]-[hairpins]. We tested a few alternative designs with the same underlying idea— different numbers of hairpins from the sgRNA scaffold, and an entirely different hairpin sequence (a Csy4 recognition sequence) were added to the pegRNA's 3' end. The design with 2 hairpins from the sgRNA scaffold boosted the overall editing efficiency by more than 2x (Figure 5.2a). Encouraged by this result, we tested additional hairpin designs. Manually designed sequences that were predicted to fold into 2, 3, or 4 consecutive hairpins on the 3' end of the pegRNA were tested (Figure 5.2b). Ultimately, the 2 hairpins from the sgRNA scaffold exhibited superior performance to all other options. Concurrent with this work, Nelson et al. created engineered pegRNAs (epegRNAs)[14], which are similar to the pegRNAs accessorized with hairpins in this chapter. The meticulous engineering of epegRNAs shed light on the mechanism of 3' modifications boosting editing rates—the structured motifs stabilize the pegRNAs by blocking exoribonucleases from degrading the 3' extensions. Nelson et al. found pseudoknots to be the most efficacious sequences for this purpose, so we tried implementing them in later designs.

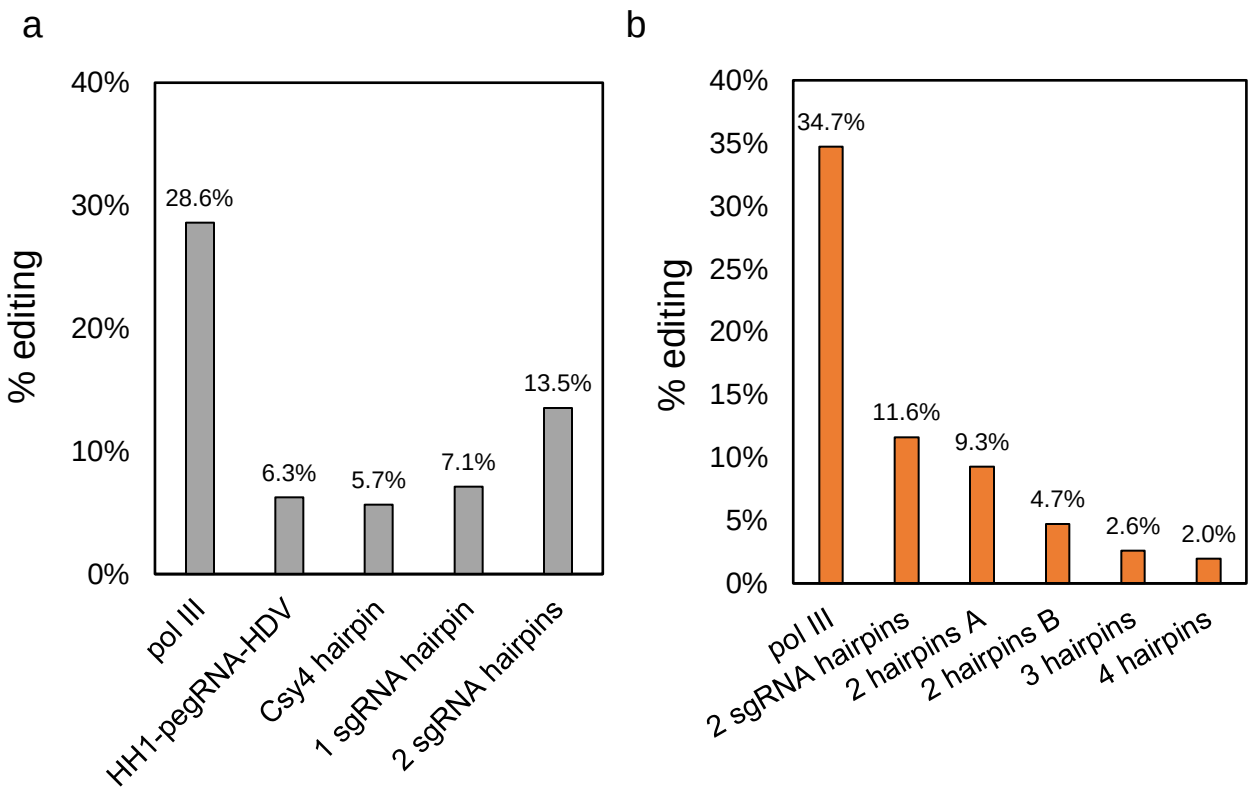


Figure 5.2. In an endeavor to mimic the structure of an sgRNA's 3' end at the extreme 3' end of a pegRNA, various hairpin sequences were appended to the 3' extension of a 6xHis-inserting pegRNA. **a,** 1 or 2 hairpins with the endogenous sequence of the sgRNA's scaffold were added to the pegRNA's 3' end. Alternatively, an unrelated hairpin sequence (a Csy4 recognition motif) was appended to the pegRNA. **b,** 2 manually designed hairpin sequences predicted to have the same overall structure as the top-performing "2 sgRNA hairpins" design from panel a were tested. These are denoted as "2 hairpins A" and "2 hairpins B." Similarly, manually designed sequences predicted to fold into 3 or 4 hairpins were added to the 3' ends of a 6xHis-inserting pegRNA.

5.3.3 Propagating pairs of pegRNAs for peCHYRON were expressed via RNAP II, with various modifications to the 3' extensions and expression cassette

When we procured two pairs of propagating pegRNAs with comparable efficiencies to each other (A-B_4 and A-B_7, from the high-throughput screen described in Chapter 4 and Supplementary Figure 4.2), we substituted them into the RNAP II expression cassettes with self-cleaving ribozymes on either side. We tested several new approaches, in parallel, to improve editing rates. Variations of the pegRNAs, themselves, included combinations of the following: optimized scaffold sequences with longer hairpins to interact with Cas9 (opt_scaff)[15], 2 hairpins from the sgRNA scaffold on the end of the 3' extension, a triplex RNA sequence on

the end of the 3' extension, and a pseudoknot (mpknot) on the end of the 3' extension. The triplex and mpknot sequences were tested because we learned from Nelson et al. that protecting the 3' extension from degradation is highly beneficial to editing efficiency. The pseudoknot is the exact sequence characterized by Nelson et al.[14], and the triplex sequence has been previously reported to protect RNA from degradation[16]. Additionally, the expression cassette was modified by removing its poly(A) signal. We hypothesized this could aid ribozyme folding by allowing the transcript to linger in the cytoplasm for a longer duration, giving the ribozyme more time to properly fold into a catalytic conformation. With an unrelated pegRNA, we found in prior experiments that removing the poly(A) signal yielded significant increases in editing efficiency (data not shown), so we excluded the poly(A) signal from most of the tested constructs in this panel. Also, with the same underlying idea as removing the poly(A) signal, we attempted to express the ribozyme-flanked pegRNAs from within introns that splice into lariat-like configurations and remain in the cytoplasm (HSV-1 LAT intron and sno-lncRNA2 intron). Nissim et al. showed this was an effective strategy for sgRNA expression from pol II transcripts[1]. After transfecting HEK293Ts that harbored a peCHYRON recording locus (sequence "A") at the genomic locus HEK3 with PE2* and the various pairs of propagating pegRNAs, the ribozyme approach appeared to be impractical. For both sets of propagating pegRNAs, the A->B editing step yielded a maximum efficiency of 7% over 3 days (Figure 5.3a-b). In comparison, the same pegRNAs expressed with an RNAP III-driven promoter typically edit with 35-40% efficiency. After testing these numerous designs for optimization, these RNAP II-expressed pegRNAs are only ~20% as efficient as their RNAP III-expressed counterparts. The B->A editing steps are inefficient even when expressed *via* RNAP III (~6% editing for B_4->A and ~10% editing for B_7-

>A), and the best designs with RNAP II expression are only half as efficient as RNAP III (<5% editing over 3 days).

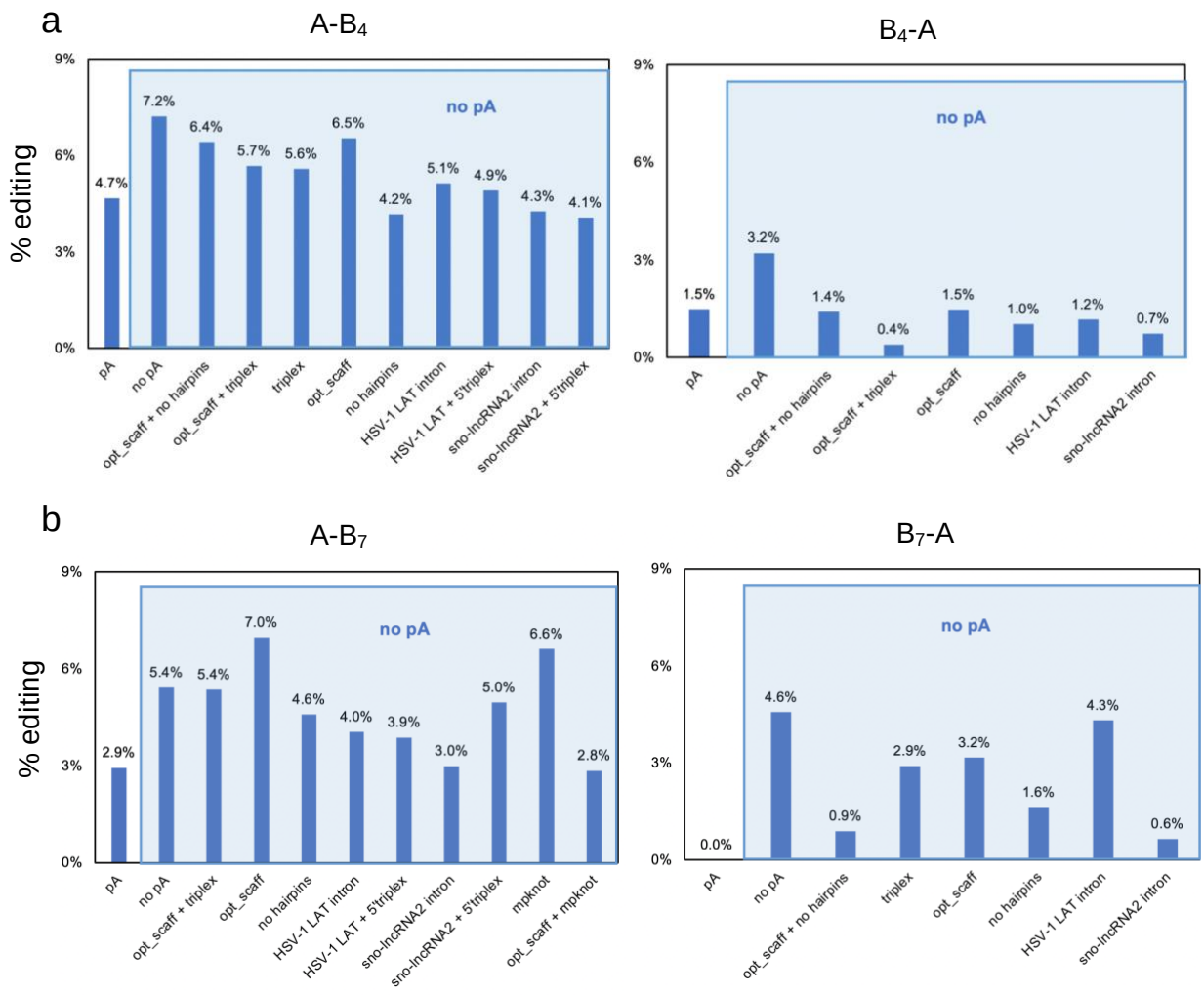


Figure 5.3. Two candidate pairs of propagating pegRNAs for peCHYRON were tested in the ribozyme-flanked RNAP II-driven cassette, with various modifications designed to boost editing efficiency. **a,** In the A-B_4 and B_4-A pair of pegRNAs, A is a 6xHis-based sequence, and B_4 is a random sequence that was enriched in the high-throughput screen for efficient editing. Every construct inside the blue-shaded box had no polyadenylation signal in the expression cassette. The descriptors along the x-axis indicate the additional modifications added to the extreme 3' end of the pegRNA, or to the expression cassette (*e.g.*, placing the ribozyme-flanked pegRNA within an intron). HEK29T cells possessing an A sequence at site3 were co-transfected with PE2* and the indicated pair of pegRNAs. After 3 days of editing, the cells were sequenced with Illumina. **b,** A second pair of propagating pegRNAs, wherein B_7 is a different enriched sequence from the high-throughput screen, was tested in the same manner as in panel a.

5.3.4 Csy4 is an efficient option for cleaving pegRNAs from RNAP II transcripts

Given the low efficiency of ribozyme-flanked pegRNAs, we forfeited our original aim of achieving RNAP II-driven pegRNA expression without the need for accessory proteins to process the mRNA. Csy4 is a CRISPR-associated endoribonuclease that processes pre-crRNAs in *Pseudomonas aeruginosa*[17]. It recognizes a specific 20 nt sequence that forms a hairpin in the RNA, then cleaves it at a predictable site. Nissim et al. established the use of Csy4 for expressing sgRNAs from RNAP II transcripts[1]. Our approach to adapting Csy4 for our peCHYRON pegRNAs was the same as for the ribozyme-flanked pegRNAs. We tested several different combinations of 3' extension modifications, pegRNA scaffold modifications, and expression cassette variations. The best combination of features was seen in the pair A-B_4, with

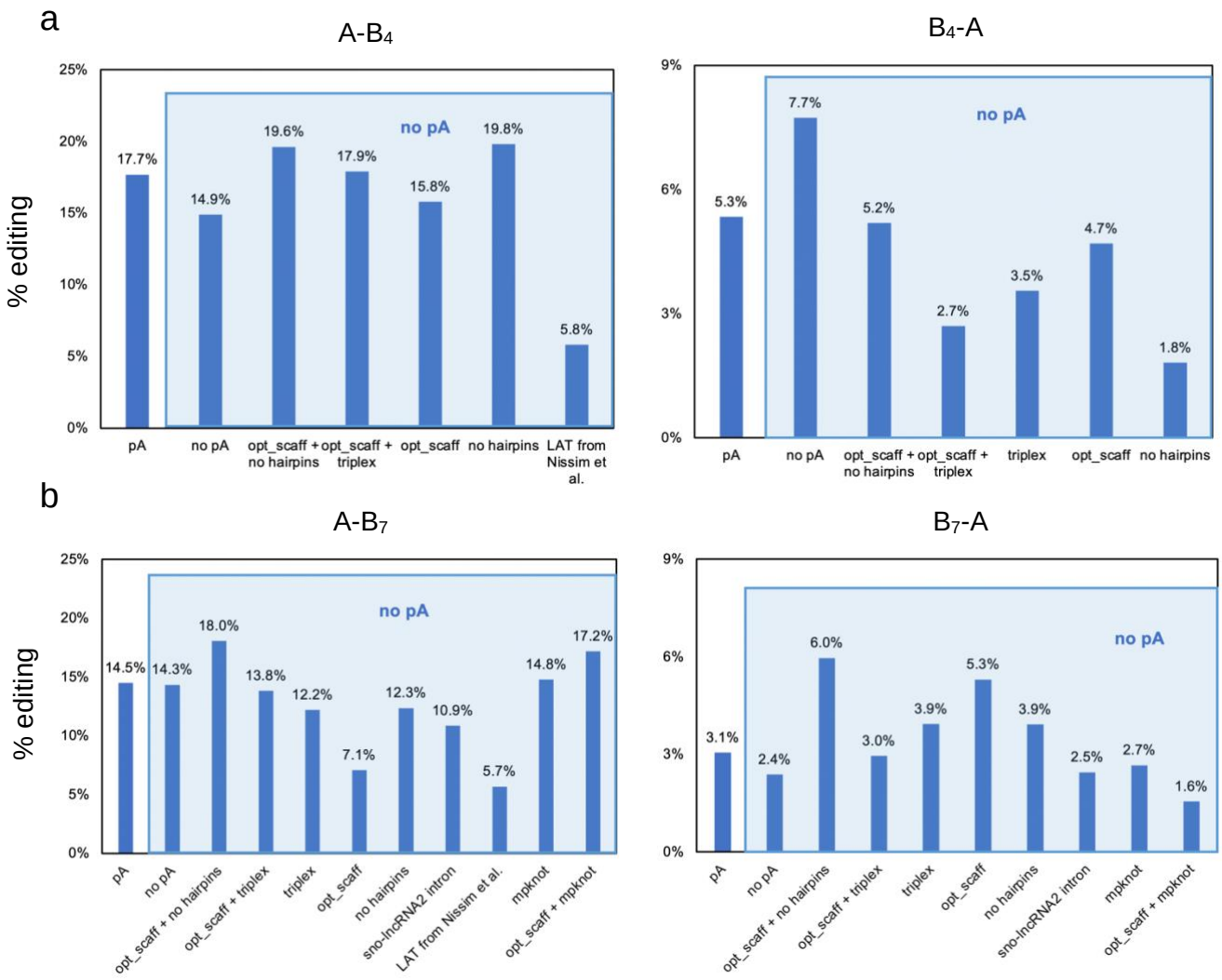


Figure 5.4. Two candidate pairs of propagating pegRNAs for peCHYRON were flanked with Csy4 recognition sequences and tested in an RNAP II-driven cassette, with various modifications designed to maximize editing efficiency. **a,** In the A-B_4 and B_4-A pair of pegRNAs, A is a 6xHis-based sequence, and B_4 is a random sequence that was enriched in the high-throughput screen for efficient editing. Every construct inside the blue-shaded box had no polyadenylation signal in the expression cassette. The descriptors along the x-axis indicate the additional modifications added to the extreme 3' end of the pegRNA, or to the expression cassette (*e.g.*, placing the ribozyme-flanked pegRNA within an intron). HEK29T cells possessing an A sequence at site3 were co-transfected with Csy4, PE2* and the indicated pair of pegRNAs. After 3 days of editing, the cells were sequenced with Illumina. **b,** A second pair of propagating pegRNAs, wherein B_7 is a different enriched sequence from the high-throughput screen, was tested in the same manner as in panel a.

approximately 20% editing efficiency for the A->B step (compared to 35% for its RNAP III-expressed equivalent) and approximately 7.5% efficiency for the B->A step (this is as good as the RNAP III-expressed version).

5.3.5 A panel of inducible pegRNAs were established for peCHYRON

The best-performing cassettes for RNAP II-expressed peCHYRON pegRNAs had their constitutive CMV promoters replaced with inducible promoters. We selected hypoxia-, Wnt- and doxycycline- inducible promoters because there are existing synthetic promoters for each of these that have been well characterized, and hypoxia and Wnt are known to play pivotal roles in various tissue development processes. For instance, Wnt is credited with helping to orchestrate cell fate determination, motility, polarity, primary axis formation, organogenesis, and stem cell renewal[18]. Defective Wnt signaling is implicated in many human pathologies, such as skeletal defects, birth defects (including the most common birth defect affecting the central nervous system: spina bifida), and cancers of the breast, colon, and skin. Hypoxia is known to influence cell fate during many different developmental processes, including solid tumor progression[19]. In hypoxic conditions, the transcription factor HIF1α is stabilized, which is correlated with activation of the "angiogenic switch," in which vascularization of a tumor is increased[20]. Although doxycycline doesn't play a natural role in tissue development, we imagined having an exogenously inducible recorder could be useful. To illustrate this point: we envisioned eventually establishing an animal model that records biologically interesting signals acting on its tissues,

such as hypoxia and Wnt, but is periodically exposed to pulses of induction with the doxycycline recorder at known times. The resulting dox-indicating signatures in the records would allow the records in the peCHYRON locus to be mapped to real time stamps.

We directly substituted inducible promoters into the top performing A->B_4 and B_4->A expression constructs with Csy4 sites flanking the pegRNAs. The 4x-HRE-YB_TATA promoter[21] and 7xTCF-LEF[8] promoters were used to achieve hypoxia- and Wnt-dependent expression. For doxycycline-responsive expression, we used the TRE3G promoter and the Tet-On® 3G reverse tetracycline transactivator[22]. The inducible pegRNA genes for the A->B step of peCHYRON propagation were genomically integrated via piggyBAC transposition into HEK293Ts harboring a peCHYRON recording locus (sequence “A”). For one set of samples, prime editor (PE2*) was also genomically integrated *via* piggyBAC at the same time, while another set of samples was instead transiently transfected with PE2*. Successfully transformed cells were selected with puromycin. Populations were exposed to a chemical hypoxia mimic (DMOG), a chemical Wnt mimic (LiCl), doxycycline, or no inducer, and sequenced *via* Illumina after 3 days. Because this experiment was conducted in single replicate, the results for both stable PE2* and transient PE2* expression are shown, so the overall trends can be compared for similar experiments. Both versions of the experiment produced an increase in prime editing in the presence of inducer, for all 3 recorders (Figure 5.5).

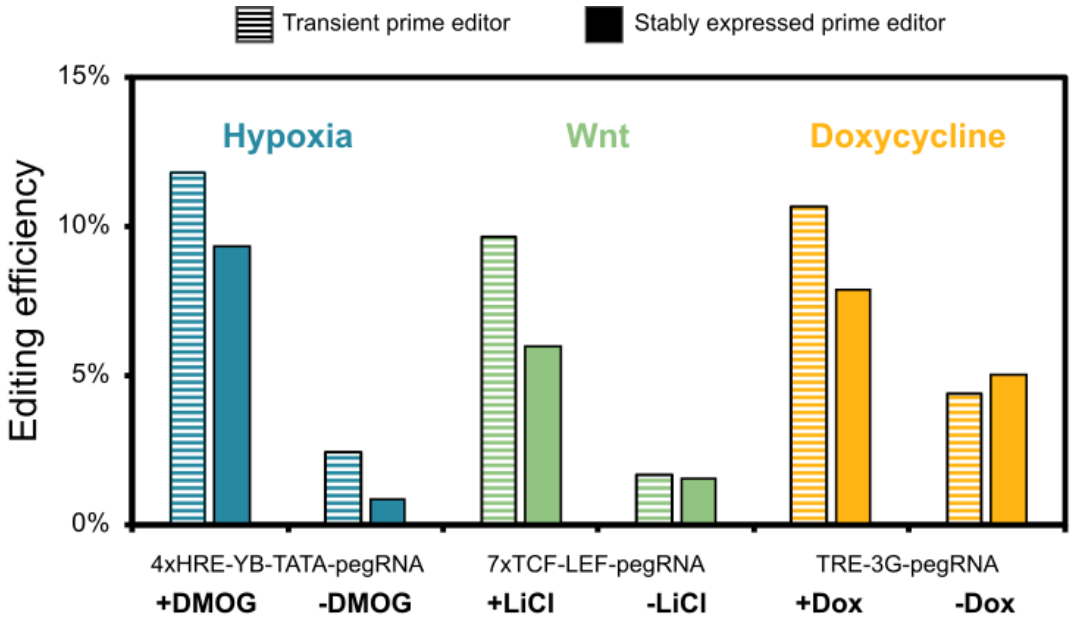


Figure 5.5. RNAP II-transcribed pegRNAs were used to record three different chemical stimuli *via* prime editing. After optimization experiments with a constitutive RNAP II promoter (CMVp), the best-performing A->B pegRNA construct was retrofitted with inducible Pol II promoters that could sense and respond to hypoxia, Wnt, and doxycycline. The inducible pegRNA constructs were genomically integrated with piggyBAC transposase, and prime editor was either transiently transfected or genomically integrated into the samples. After 3 days of induction with a chemical hypoxia mimic (DMOG), a chemical Wnt mimic (Wnt), or doxycycline, the populations were sequenced with Illumina.

5.4 Discussion

To bring this inducible pegRNA project to its logical conclusion, two main objectives need to be met: 1) the phenomenon of pegRNA "cannibalization," which was discussed in Chapter 4, needs to be ameliorated, and 2) a demonstration of multiplexed signal recording needs to be completed. Mitigating pegRNA cannibalization should first be attempted with the same approach that proved effective for the RNAP III-expressed pegRNAs; that is, constitutively nicking the same strand that receives the unwanted edit. The architecture of an RNAP II-expressed gene is distinct from an RNAP III-expressed gene in that the Pol II gene has a 3' UTR, a long (~60 bp) transcription termination signal, and imprecise locus of transcription termination. In contrast, the RNAP III-transcribed gene is only comprised of the pegRNA's coding sequence, and

transcription is terminated abruptly by a 7-bp homopolymeric tract of thymidines immediately downstream of the pegRNA gene. The cannibalization edit occurs within the pegRNA gene, and in the case of the RNAP III-transcribed pegRNAs, we found nicking <50 bp away from the site of cannibalization to be most effective at curbing this undesired edit. However, in the Pol II construct, nicking this close to the cannibalization edit may be problematic because it would necessitate nicking within the RNAP II-transcribed gene. This may interfere with expression due to the RNA polymerase colliding with the nickase Cas9. On the other hand, nicking downstream of the transcription termination signal may be too distant from the cannibalization edit to have an effect. Nonetheless, I recommend attempting the nicking approach first. If nicking is unsuccessful for this setup, another appealing option is to encode an intron within the pegRNA genes, such that the prime editing target site required for cannibalization is disrupted by the intron, but spliced out of the transcribed pegRNAs. This approach will require more involved engineering, however. Once cannibalization is fixed, the effective strategy for blocking cannibalization should be invoked at the same time pulsed expression of the pegRNAs is recorded with induction of hypoxia, Wnt, and doxycycline.

For future development of RNAP II-expressed pegRNAs, it is important to consider the findings of Nelson et al.[14] Specifically, while developing epegRNAs, they showed that the 5' ends of pegRNAs are shielded from exonucleases by their association with Cas9, but the 3' ends, which protrude from the protein, are frequently destroyed by exonucleases. This leads to a collection of pegRNAs that are competent for target site searching and binding *via* their intact 5' ends, but incapable of installing any edits due to their decimated 3' extensions; therefore, these pegRNAs can cause prime editor to block the target site and drastically reduce editing rates. Avoiding this phenomenon of pegRNA 3' extension degradation will be important when seeking maximal efficiency with peCHYRON; therefore, it is recommended to proceed with designs that protect the 3' end. The leading designs with Csy4 likely provide sufficient protection, since the Csy4 protein remains bound to the 3' end of the pegRNA after cleavage. If,

however, the Csy4 scheme is replaced in the future, care should be taken to guard the pegRNAs' 3' ends.

5.5. References

1. Nissim, L., Perli, S. D., Fridkin, A., Perez-Pinera, P. & Lu, T. K. Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells. *Mol Cell* **54**, 698–710 (2014).
2. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science (1979)* **353**, aag0511 (2016).
3. Kallunki, T., Barisic, M., Jäättelä, M. & Liu, B. How to choose the right inducible gene expression system for mammalian studies? *Cells* **8**, 796 (2019).
4. Romanova, N. & Noll, T. Engineered and natural promoters and chromatin-modifying elements for recombinant protein expression in CHO cells. *Biotechnol J* **13**, 1700232 (2018).
5. Hirose, Y. & Manley, J. L. RNA polymerase II and the integration of nuclear events. *Genes Dev* **14**, 1415–1429 (2000).
6. Liu, P. *et al.* Improved prime editors enable pathogenic allele correction and cancer modelling in adult mice. *Nat Commun* **12**, 2121 (2021).
7. Ede, C., Chen, X., Lin, M.-Y. & Chen, Y. Y. Quantitative analyses of core promoters enable precise engineering of regulated gene expression in mammalian cells. *ACS Synth Biol* **5**, 395–404 (2016).
8. Veeman, M. T., Slusarski, D. C., Kaykas, A., Louie, S. H. & Moon, R. T. Zebrafish prickle, a modulator of noncanonical Wnt/Fz signaling, regulates gastrulation movements. *Current biology* **13**, 680–685 (2003).

9. Das, A. T., Zhou, X., Metz, S. W., Vink, M. A. & Berkhout, B. Selecting the optimal Tet-On system for doxycycline-inducible gene expression in transiently transfected and stably transduced mammalian cells. *Biotechnol J* **11**, 71–79 (2016).
10. Loveless, T. B. *et al.* Lineage tracing and analog recording in mammalian cells by single-site DNA writing. *Nat Chem Biol* **17**, 739–747 (2021).
11. Fonseca, J. P. *et al.* A toolkit for rapid modular construction of biological circuits in mammalian cells. *ACS Synth Biol* **8**, 2593–2606 (2019).
12. Mahon, M. J. Vectors bicistronically linking a gene of interest to the SV40 large T antigen in combination with the SV40 origin of replication enhance transient protein expression and luciferase reporter activity. *Biotechniques* **51**, 119–126 (2011).
13. Anzalone, A. V *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
14. Nelson, J. W. *et al.* Engineered pegRNAs improve prime editing efficiency. *Nat Biotechnol* **40**, 402–410 (2022).
15. Dang, Y. *et al.* Optimizing sgRNA structure to improve CRISPR-Cas9 knockout efficiency. *Genome Biol* **16**, 1–10 (2015).
16. Wilusz, J. E. *et al.* A triple helix stabilizes the 3′ ends of long noncoding RNAs that lack poly (A) tails. *Genes Dev* **26**, 2392–2407 (2012).
17. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence-and structure-specific RNA processing by a CRISPR endonuclease. *Science (1979)* **329**, 1355–1358 (2010).
18. Komiya, Y. & Habas, R. Wnt signal transduction pathways. *Organogenesis* **4**, 68–75 (2008).
19. Stampone, E. *et al.* Genome editing and cancer therapy: handling the hypoxia-responsive pathway as a promising strategy. *Cellular and Molecular Life Sciences* **80**, 220 (2023).

20. Giles, R. H. *et al.* Interplay between VHL/HIF1α and Wnt/β-catenin pathways during colorectal tumorigenesis. *Oncogene* **25**, 3065–3070 (2006).

21. Ede, C., Chen, X., Lin, M.-Y. & Chen, Y. Y. Quantitative analyses of core promoters enable precise engineering of regulated gene expression in mammalian cells. *ACS Synth Biol* **5**, 395–404 (2016).

22. Das, A. T., Zhou, X., Metz, S. W., Vink, M. A. & Berkhout, B. Selecting the optimal Tet-On system for doxycycline-inducible gene expression in transiently transfected and stably transduced mammalian cells. *Biotechnol J* **11**, 71–79 (2016).

CHAPTER 6: Conclusions, lessons learned, and future of our DNA recorders

6.1 Summary

In this dissertation, I worked together with members of the Liu lab, especially Theresa Loveless, who was my co-first author on the peCHYRON paper and the first author of the CHYRON paper, to develop two distinct *in vivo* DNA recorders. The first recorder is titled Cell HistorY Recording by Ordered iNsertion (CHYRON). It combines the activities of a homing guide RNA (hgRNA), Cas9 nuclease, and template-independent polymerase called TdT to iteratively write insertion mutations at a neutral locus in the host cells' genomes. The second DNA recorder, prime editing CHYRON (peCHYRON) achieves the same overall behavior as CHYRON (continuous, iterative accumulation of insertions), but it uses prime editor and an array of prime editing guide RNAs (pegRNAs) to propagate the recording locus. Both systems are capable of both cell lineage tracing and analog signal recording. My primary contributions to the CHYRON technology were: attempting to replace Cas9 with Cpf1, which I maintain is a promising avenue of development for the future, establishing a hypoxia-responsive version of CHYRON that only logs insertions when the cells detect hypoxia, and engineering TdT to have altered nucleotide insertion preferences. The engineered TdTs improved the diversity of barcodes for cell lineage tracing and unlocked a new dual-channel signal recording modality, which was demonstrated for recording of hypoxia and Wnt in populations of HEK293Ts. Toward the establishment of peCHYRON, we engineered pegRNAs that can encode information-rich "signature mutations" (which provide the sequence diversity needed for cell lineage tracing, or record the identities of the induced pegRNAs for multiplexed signal recording) *via* a combination of rational design and high-throughput screening *in vivo*. The rate and fidelity of insertion accumulation over time were characterized in a 67-day timecourse, showing little to no reduction in efficiency and minimal corruption of records. This system was used for cell lineage tracing *in vitro* and a proof-of-

principle multiplexed signal recording experiment that used synthetic inducible RNAP III-transcribed pegRNAs. We also worked toward developing RNAP II-transcribed pegRNAs, which will allow a vast selection of biologically interesting signals to be recorded with peCHYRON.

6.2 Comparative merits of CHYRON and peCHYRON

In their present incarnations, CHYRON and peCHYRON have complementary strengths and weaknesses that should be considered when choosing which applications are well-suited to each recorder. For cell lines that can't be easily transformed, CHYRON may be the best choice because it requires few transgenes to be delivered per cell (only the hgRNA + Cas9-T2A-TdT), whereas peCHYRON needs the large PE gene to be introduced, plus many pegRNA genes to achieve diverse editing outcomes (*e.g.*, for lineage tracing, we used ~40 pegRNA genes). However, if the goal of a project is to record one environmental signal, transgene delivery for peCHYRON may be feasible, since only one pegRNA cassette is needed per signal. Another important consideration is the efficiency of editing—CHYRON edits much faster than peCHYRON (~80% of loci are edited in 3 days, compared to ~30% and ~3% in 3 days for the A->B and B->A steps of peCHYRON). The fidelity of records is a distinctive strength of peCHYRON, as it only produces indels ~1% of the time it edits. In contrast, CHYRON creates a deletion ~20% of the time. Related to the fidelity of editing, peCHYRON has no upper limit on the duration of recording it can perform, since each correct edit renews the target site for the next round of editing. CHYRON, on the other hand, ceases editing after approximately 3 rounds of insertion mutagenesis. Another advantage of peCHYRON is that it can theoretically record up to 64 signals in multiplex, whereas CHYRON can only record 2 signals at a time. However, it should be noted multiplexed signal recording the RNAP II-expressed pegRNAs has not yet been achieved with peCHYRON. Also, peCHYRON inherently installs punctuation (the "propagator sequences") between each edit, which improves the accuracy of analog signal recording and simplifies the analysis of lineage relationships. A final consideration when weighing the two

recorders is that peCHYRON relies solely on prime editor for mutagenesis, which is less likely to disrupt cell biology because it inflicts little/no off-target editing on the cells of interest[1]. CHYRON, on the other hand, poses a greater risk of interfering with the biological processes in question, as Cas9 is known to incur off-target edits in the host cells' genome[2]. TdT may also interfere with cell health, as it is recruited non-specifically to DSBs and could therefore mutagenize loci throughout the genome when naturally occurring DSBs arise. However, it is worth noting Cas9 and TdT have been successfully used to study hematopoiesis in the DARLIN mouse line[3], indicating that the health risks of these enzymes may be tolerable in animal models. Also, if future implementations of CHYRON reveal practical issues due to Cas9 and/or TdT mutating the host genome in an undesirable manner, measures may be taken to improve the safety to host cells (see below).

6.3 Future directions for CHYRON

Replacing CHYRON's Cas9 and hgRNA with a nuclease that can invoke repeated rounds of editing without a self-targeting guide would hold great promise in improving the durability of recording. This is because the guide RNA wouldn't change in length over time (which is known to affect editing efficiency over time), and deletions in the recording locus wouldn't necessarily terminate editing. Cpf1 is an appealing candidate for a nuclease that could fill this role because its seed region is distal to its cut site[4]; therefore, mutations that accrue at the cut site would have a small impact on editing efficiency because they wouldn't arise in the seed region. However, Cpf1 creates DSBs with a 5' overhang, which is a strongly disfavored substrate for TdT addition[5]. A relatively simple approach to achieving efficient collaboration between Cpf1 and TdT would be to test a TdT-Klenow fragment fusion. The Klenow fragment could fill in the recessed 3' end, resulting in a blunt ended DSB, which TdT can extend moderately efficiently. Because this approach is relatively facile to implement and small in scale (testing several different linkers between TdT and the Klenow fragment would be advisable, but even this would

result in only a handful of designs to be tested), I would recommend pursuing this line of experimentation first. If the Klenow Fragment-TdT fusions are not efficient, an alternative strategy is to forego the Klenow Fragment, and instead engineer TdT to efficiently add onto recessed 3' ends. Ideally, TdT would fill in the recessed 3' end in template-dependent mode, then once the DNA ends have become blunt, switch to template-independent extension. The massively parallel TdT screen described in Chapter 3 could facilitate a large library-scale approach for this. Loop1, which sterically blocks double-stranded DNA from entering TdT's active cavity[5], would be a logical choice for saturation mutagenesis in this screen. Encouragingly, others have shown that removing Loop1 evokes template-dependent activity from TdT, and splicing Loop1 onto a closely related template-dependent polymerase (pol mu) causes it to perform template-independent synthesis [6]. Furthermore, we found in our high-throughput screen of TdT variants that residues in Loop1 have a decisive impact on the nucleotide preferences of TdT, supporting the idea that this loop plays a role in recruiting nucleotides into the active site in the absence of a template strand to base pair with. Therefore, modifying Loop1 could be a fruitful approach to finding a flexible phenotype that can perform template-dependent synthesis (to fill in the recessed 3' end), then also perform template-independent insertion mutagenesis. Engineering Cpf1 to make blunt cuts could also effective, and the same high-throughput assay that was implemented for TdT engineering could be used to screen for effective Cpf1 variants. An additional engineering thrust that should be pursued for CHYRON is recruiting TdT specifically to the recording locus. This could potentially reduce the rate of deletions by increasing the local concentration of TdT at the double strand break, which would improve the durability of recording (by preventing the PAM in the hgRNA gene from being deleted) and reduce the rate of record corruption. Also, recruiting TdT to the recording locus could provide a strategy to hedge against negative health effects to host organisms, as it would allow TdT to be expressed at a very low level in the cell and still achieve high local concentration at the recording locus.

6.4 Future directions for peCHYRON

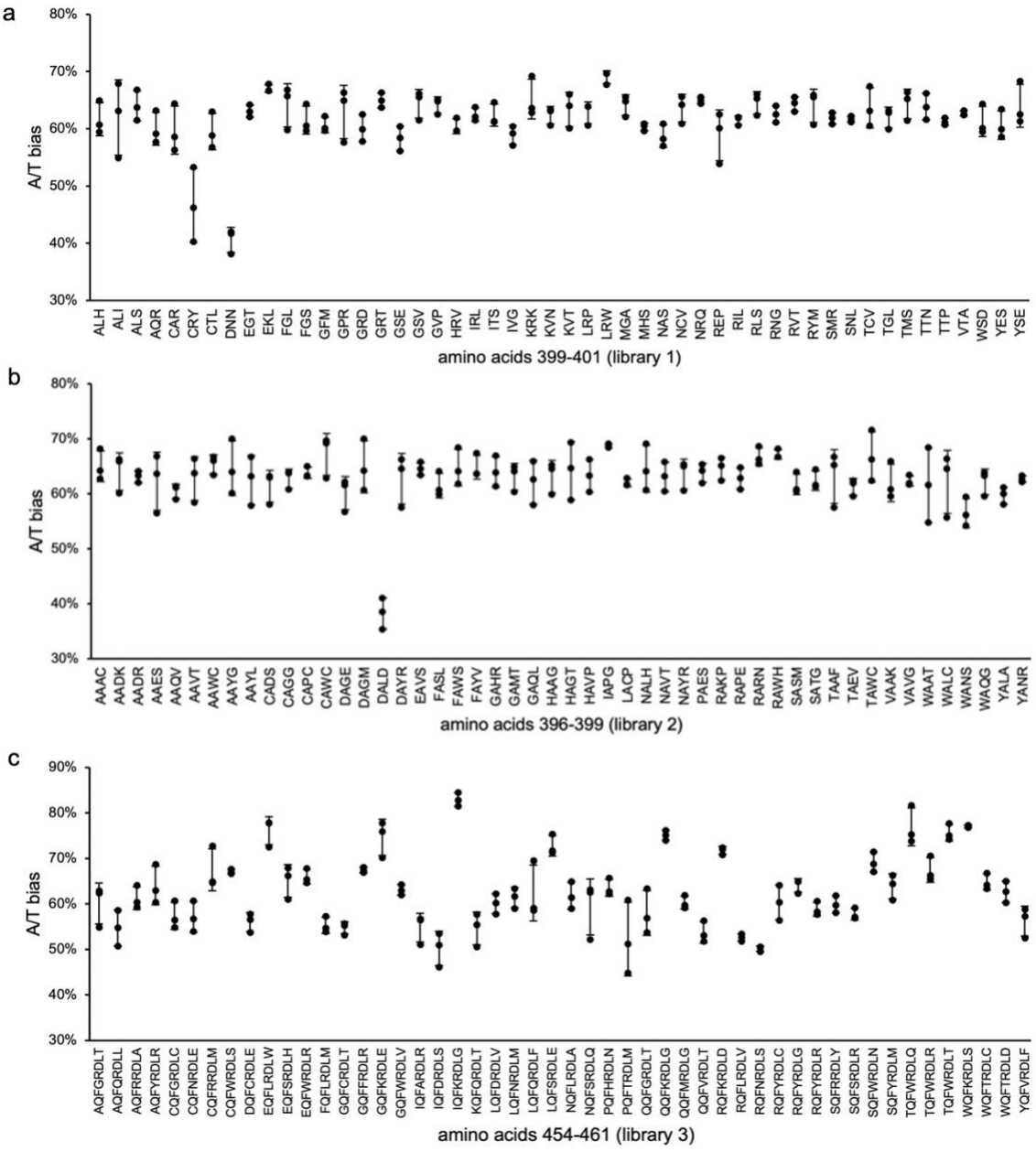
Improving the efficiency of editing with peCHYRON is the highest priority at the moment. pegRNA design tools that use AI to predict efficient pegRNA sequences[7–10] could allow small libraries of A->B and B->A pegRNAs that are anticipated to efficiently propagate the peCHYRON locus to be specifically synthesized and screened. In addition to pegRNA engineering, the prime editor protein could be substituted for a new evolved variant that is specialized for long insertion sequences (demonstrated for 20 nt insertions in one published study)[11]. Once editing efficiency is increased, long insertions at the recording locus will become more frequent. This will be a boon to peCHYRON's ability to record information. For multiplexed signal recording, it will only become possible to record many signals in parallel if edits are logged very frequently, otherwise it's impossible to capture all of the pegRNAs that are induced contemporaneously. Furthermore, individual cells need to make multiple edits in order to store temporal information (*i.e.*, the ordered nature of records doesn't matter if multiple insertions aren't generated at a single target site). For lineage tracing, increasing the rate of insertion mutagenesis will be necessary to approach the resolution of single-cell lineages. At least one edit would need to occur per cell division if individual cells' relationships are all to be comprehensively mapped. However, we note that the repetitive nature of the recording locus may create locus instability or difficulty PCR-amplifying the locus for NGS prep. If this issue comes to pass, alternating between 3 propagation sequences (instead of 2 in our current design) could help by reducing the periodicity of the repeats. Also, a droplet digital PCR[12], or emulsion PCR approach[13] may overcome issues due to collapsing the repetitive sequences during NGS library prep.

6.5 References

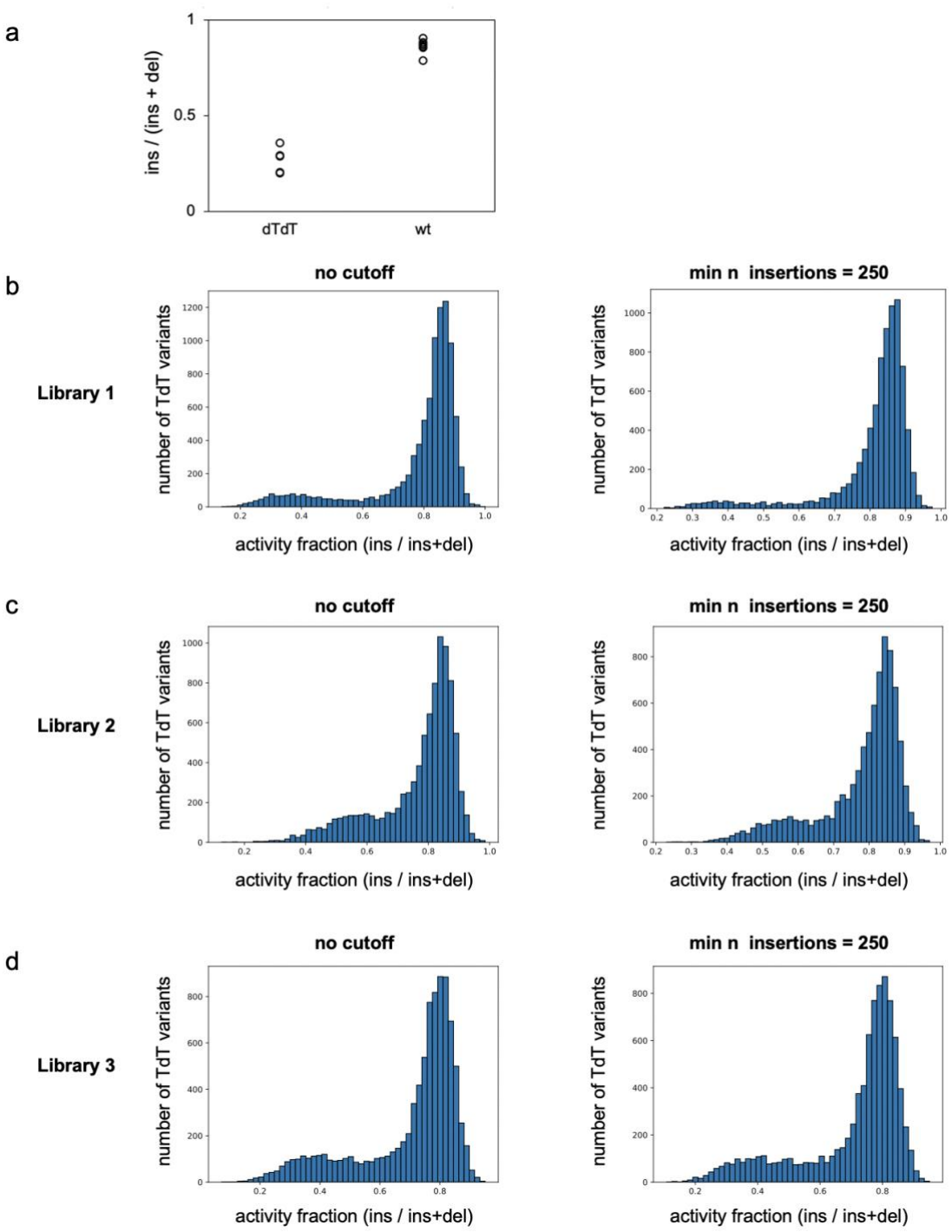
1. Liang, S.-Q. *et al.* Genome-wide profiling of prime editor off-target sites in vitro and in vivo using PE-tag. *Nat Methods* 1–10 (2023).
2. Zhang, X.-H., Tee, L. Y., Wang, X.-G., Huang, Q.-S. & Yang, S.-H. Off-target effects in CRISPR/Cas9-mediated genome engineering. *Molecular Therapy-Nucleic Acids* **4**, (2015).
3. Li, L., Bowling, S., Wang, S. & Camargo, F. 2015–A MOUSE MODEL WITH HIGH CLONAL BARCODE DIVERSITY FOR JOINT LINEAGE, TRANSCRIPTOMIC, AND EPIGENOMIC PROFILING IN SINGLE CELLS. *Exp Hematol* **124**, S44 (2023).
4. Safari, F., Zare, K., Negahdaripour, M., Barekati-Mowahed, M. & Ghasemi, Y. CRISPR Cpf1 proteins: structure, function and implications for genome editing. *Cell Biosci* **9**, 1–21 (2019).
5. Motea, E. A. & Berdis, A. J. Terminal deoxynucleotidyl transferase: the story of a misguided DNA polymerase. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1804**, 1151–1166 (2010).
6. Juarez, R., Ruiz, J. F., McElhinny, S. A. N., Ramsden, D. & Blanco, L. A specific loop in human DNA polymerase mu allows switching between creative and DNA-instructed synthesis. *Nucleic Acids Res* **34**, 4572–4582 (2006).
7. Mathis, N. *et al.* Predicting prime editing efficiency and product purity by deep learning. *Nat Biotechnol* 1–9 (2023).
8. Li, Y., Chen, J., Tsai, S. Q. & Cheng, Y. Easy-Prime: a machine learning–based prime editor design tool. *Genome Biol* **22**, 1–11 (2021).
9. Kim, H. K. *et al.* Predicting the efficiency of prime editing guide RNAs in human cells. *Nat Biotechnol* **39**, 198–206 (2021).
10. Liu, F. *et al.* Design of prime-editing guide RNAs with deep transfer learning. *Nat Mach Intell* 1–14 (2023).

11. Doman, J. L. *et al.* Phage-assisted evolution and protein engineering yield compact, efficient prime editors. *Cell* **186**, 3983–4002 (2023).
12. Hou, Y., Chen, S., Zheng, Y., Zheng, X. & Lin, J.-M. Droplet-based digital PCR (ddPCR) and its applications. *TrAC Trends in Analytical Chemistry* **158**, 116897 (2023).
13. Williams, R. *et al.* Amplification of complex gene libraries by emulsion PCR. *Nat Methods* **3**, 545–550 (2006).

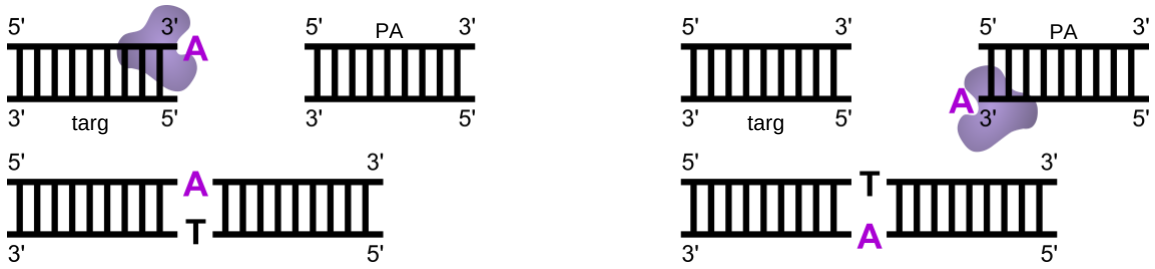
Appendix A: Supplementary information for Chapter 3



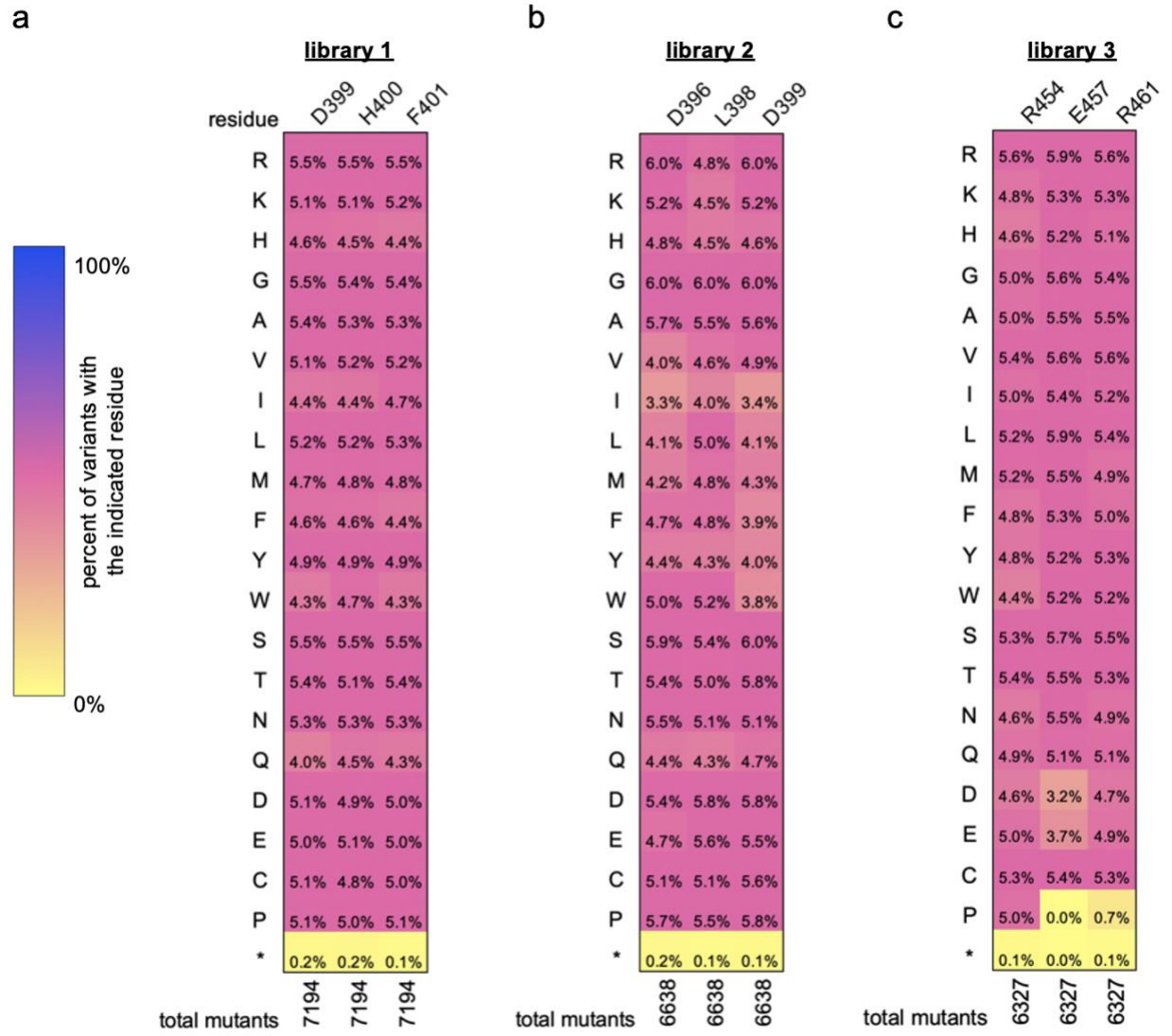
Supplementary Figure 3.1. Concordance among technical replicates of TdT variants in the high-throughput assay. **a-c,** For libraries 1-3, respectively, the apparent A/T biases of 50 random TdT variants with a minimum activity fraction of 0.7 and at least 250 insertion sequences are plotted. Each library was amplified in many replicate PCRs during Illumina prep, and 3 distinct sets of barcoding primers were used to label technical replicates. Points represent individual replicates, and error bars depict the standard deviation. Variants with close agreement among replicates are considered most likely to have reliable measurements in the HT assay.



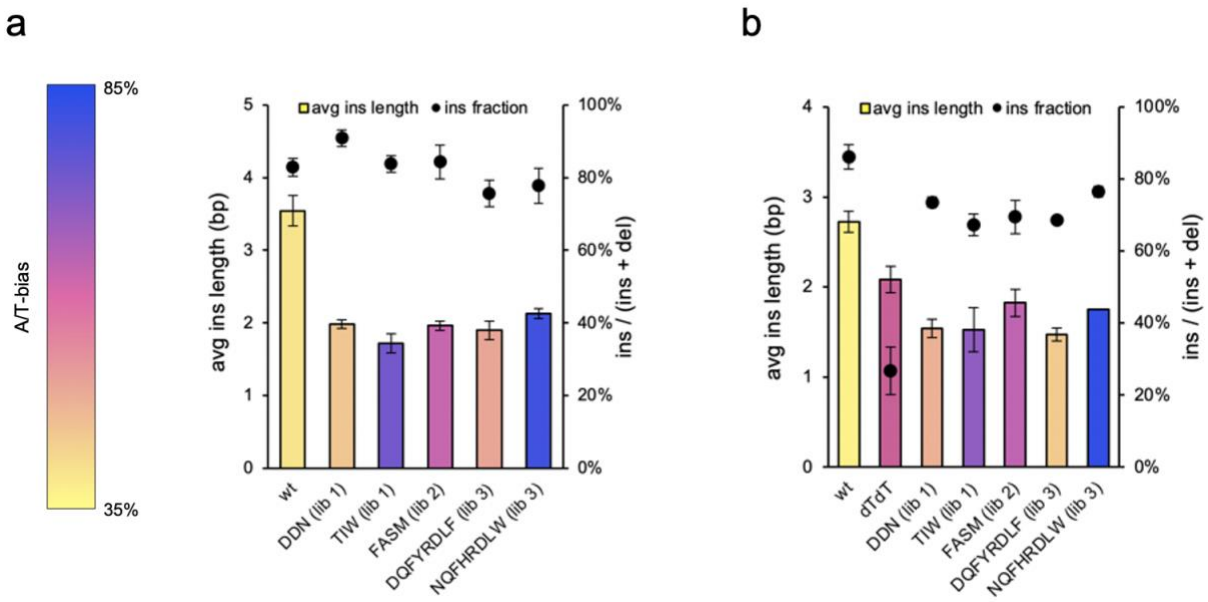
Supplementary Figure 3.2. The overwhelming majority of TdT mutants in each library are active when assayed in high-throughput. **a**, active TdT (wildtype, wt) can be discerned from a catalytically dead variant (dTdT) by computing the frequency of insertions relative to any indel (”activity fraction”). Five and eight biological replicates are shown, respectively, for dTdT and wt TdT transiently transfected into HEK293Ts with Cas9 targeting the HEK3 genomic locus. **b-d,** For libraries 1 (base of Loop1 NNKs), 2 (mid Loop1 NNKs), and 3 (alpha helix NNKs), histograms are shown for the activity fractions of the library members. The raw data for every TdT variant (no cutoff), and the data after filtering out mutants that were sampled with low statistical power (min n insertions = 250) are shown. Each TdT mutant corresponds to translated protein sequences, such that synonymous mutants in the library are counted as a single mutant.



Supplementary Figure 3.3. After Cas9 induces a double-strand break, TdT has two available primer strands to extend. This leads to ambiguity when interpreting the identities of complementary bases. If TdT uses the top strand (PAM-containing strand) as its primer, the inserted nucleotide will be read out directly during NGS. Alternatively, if TdT adds a nucleotide to the bottom strand (Cas9's target strand), the complementary base will be added to the top strand during DNA repair, and subsequently read out during NGS. Thus, TdT's preference for A/T vs. G/C can be determined from the in vivo screen, but A vs. T and C vs. G cannot.

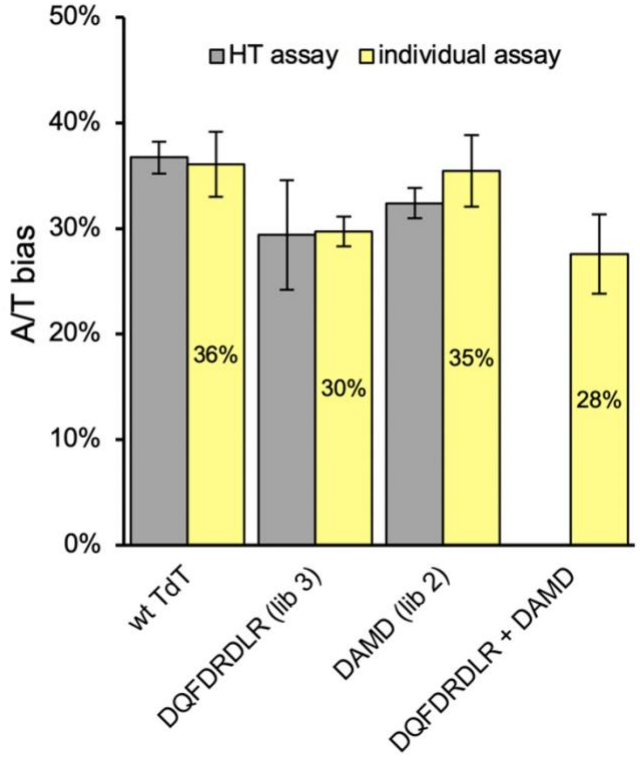


Supplementary Figure 3.4. Composition of amino acid sequences at randomized (NNK) loci in each library. Heatmaps show, for each library, what percent of TdT variants harbor a specific residue at each randomized position. **a-c,** this data shows the distribution of amino acids among all active, deeply sampled TdT variants, regardless of their A/T bias. For all 3 libraries, every TdT variant with an activity fraction greater than or equal to 0.7, and at least 250 captured insertion sequences, was considered. The composition of amino acids at each randomized site (one column per site) is shown. This data was used to normalize the compositions of amino acids among the TdT variants in distinct bias bins in Figure 3.1d.



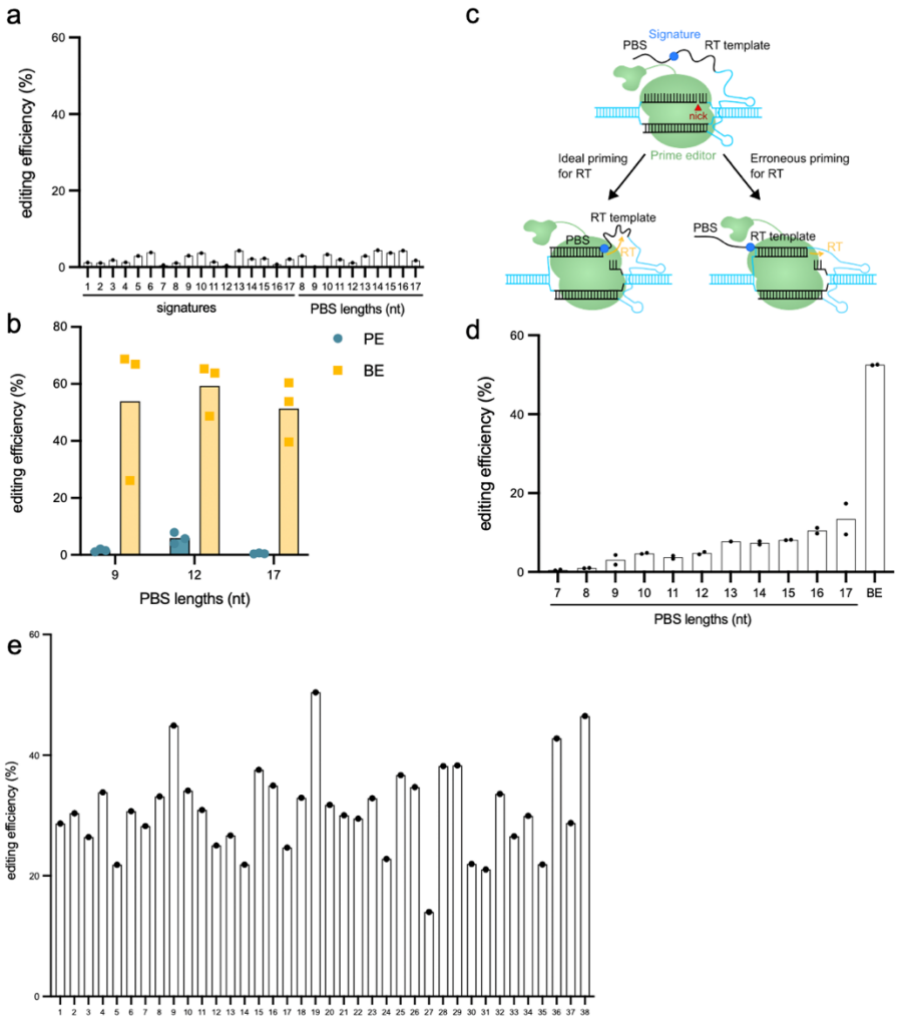
Supplementary Figure 3.5. Identification of TdT mutants that make short insertions on average but are still active when assayed *via* transient transfection. a, Data from the HT assay showing variants that generate shorter insertion mutations than wildtype TdT. Bar graphs convey the average insertion length, while points correspond to the activity fractions. Error bars represent the standard deviation of 3 technical replicates. Color coding indicates the A/T bias of each TdT variant. b, The short-inserters were individually cloned, then transfected into HEK293Ts with Cas9 and a sgRNA targeting a genomic locus. The average length of insertion mutations, A/T content of the insertion sequences, and frequency of making insertions (*i.e.*, when they are expressed, insertions accrue more often than deletions) are indicated as in a.

a



Supplementary Figure 3.6. With the intention of creating more pronounced G/C or A/T biases, mutations from separate libraries were combined. The amino acid sequences spanning all randomized residues for each variant are indicated (library 1, 399-401; library 2, 396-399; library 3, 454-461). a, A G/C-biased double mutant has a modestly stronger bias than its parent mutants. This double mutant was used for subsequent *in vitro* characterization and DNA recording experiments.

Appendix B: Supplementary information for Chapter 4



Supplementary Figure 4.1. Testing alternative pegRNA candidates for propagation. **a,** Efficiency of iterative editing with a pegRNA that can both target and insert the same 20-bp sequence (A→A) was low. The 20-bp sequence was derived from the 6xHis sequence. Constructs expressing PE2 and appropriate pegRNAs were transfected into $293T_{6xHis}$ cells. Libraries of all possible signature mutations, each with a different PBS length, were scanned. **b,** Although the efficiency of inserting the 6xHis sequence at a 6xHis target site was low, the 6xHis target site readily recruited Cas9. To ensure that the 6xHis sequence could be efficiently targeted by a pegRNA-directed Cas9, plasmids encoding selected pegRNAs from (a) and either PE2 or the C→T base editor AncBE4Max were transfected into $293T_{6xHis}$ cells, and the efficiency of a prime edit or base edit at the 6xHis sequence was measured. Points represent biological replicates. **c,** Proposed mechanism of failed prime editing with a pegRNA that both targets and inserts the same 20-bp sequence (A→A). The PBS and RT template sequence share identity, so priming can occur erroneously in the RT template instead of the PBS, yielding no net insertion. **d,** The efficiency of inserting the wt site3 sequence at a 6xHis target site was low. Here, site3 was the original sequence at the recording locus and 6xHis had been inserted (A→B) to make the $293T_{6xHis}$ cells used in this experiment. Thus, to achieve propagation, the efficiency of adding back site3 was tested (B→A). For optimization, multiple lengths of the PBS of the site3-inserting pegRNA were tested. Each sample is a library of all possible signatures with the indicated PBS length. To ensure that the 6xHis sequence could be efficiently targeted by Cas9 directed by this type of pegRNA, plasmids encoding the C→T base editor AncBE4Max and a pegRNA that targets 6xHis were co-transfected into $293T_{6xHis}$ cells, and the efficiency of a base edit in the 6xHis sequence was measured. Points represent biological replicates. **e,** Rare sequences that can efficiently insert downstream of 6xHis were identified. The top hits from the high-throughput screen shown in Figure 2c were individually cloned, and their insertion efficiencies were assayed. All experiments were conducted in 293T6xHis cells with PE2. Insertion sequences are ordered from left to right based on their enrichment scores in the original screen.

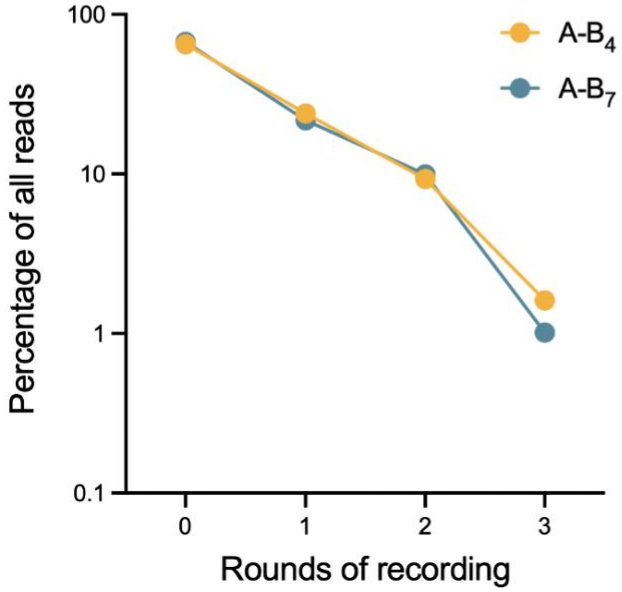
a

	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}
15	32.3%	31.2%	30.6%	22.1%	17.4%	9.3%	31.3%	24.4%	28.5%	
14	30.0%	39.5%		26.0%	21.1%	21.0%	26.7%	27.7%	29.9%	29.8%
13	25.6%	39.1%	41.8%	29.9%	12.0%		35.6%	25.6%		
12	17.5%	38.7%	41.8%	19.5%		32.1%	33.5%	29.8%	29.5%	
11	16.8%	30.3%	29.7%	37.5%	23.8%	19.1%	27.0%	22.3%	34.4%	24.2%
10	17.7%	29.8%	28.2%	29.3%	17.2%	33.0%		22.6%	28.1%	24.9%
9	8.0%	16.5%	16.6%	16.1%	10.1%	17.8%	25.1%	10.5%	19.5%	

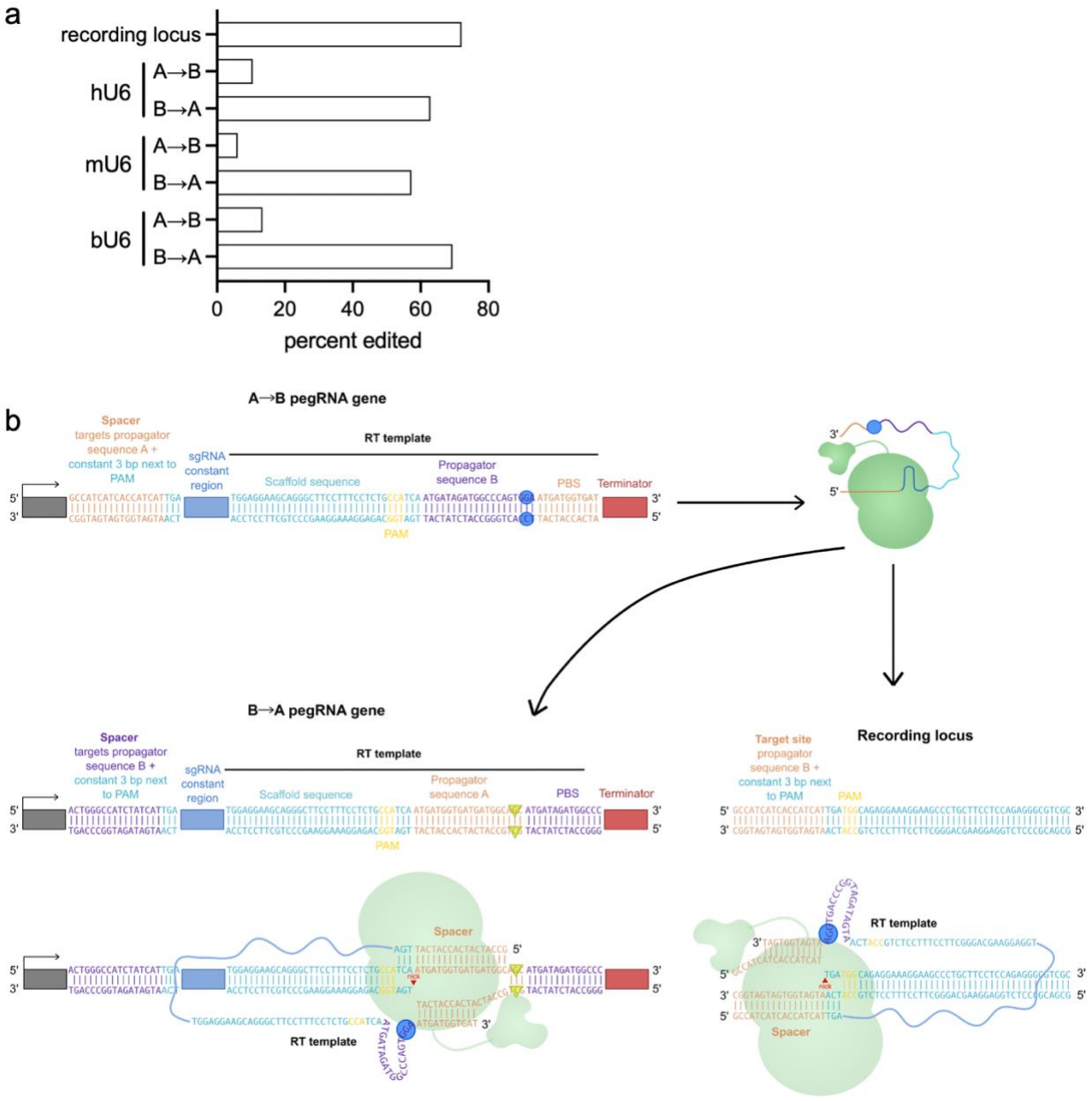
b

	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}
15	8.6%	1.0%	0.4%	17.7%	15.1%	4.2%	13.7%	5.8%	9.2%	
14	5.3%	2.4%		24.2%	21.0%	2.9%	17.8%	4.2%	7.1%	7.8%
13	6.1%	2.9%	0.8%	18.6%	19.4%	5.3%	20.3%	5.0%	12.3%	
12	5.8%	3.1%	0.4%	19.3%		5.1%	23.8%	3.5%	11.1%	5.7%
11	3.9%	3.0%	0.2%	17.7%	23.1%	4.0%	21.7%	12.0%	10.1%	6.8%
10	0.7%	2.3%	0.4%	11.1%	11.7%	2.5%		9.6%	5.8%	3.3%
9	0.3%	1.5%	0.3%	8.2%	13.1%	2.4%	6.9%	6.4%	5.7%	

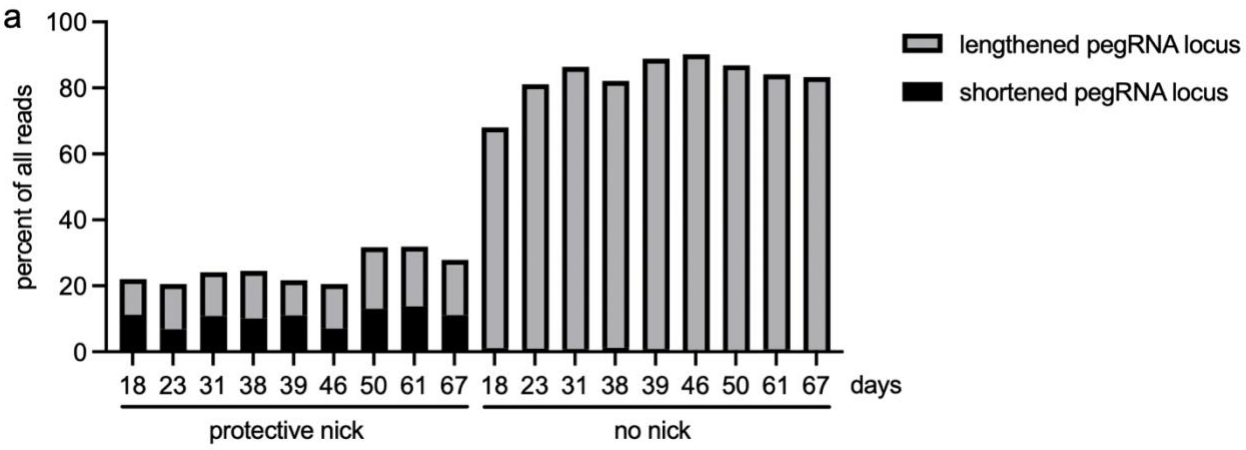
c



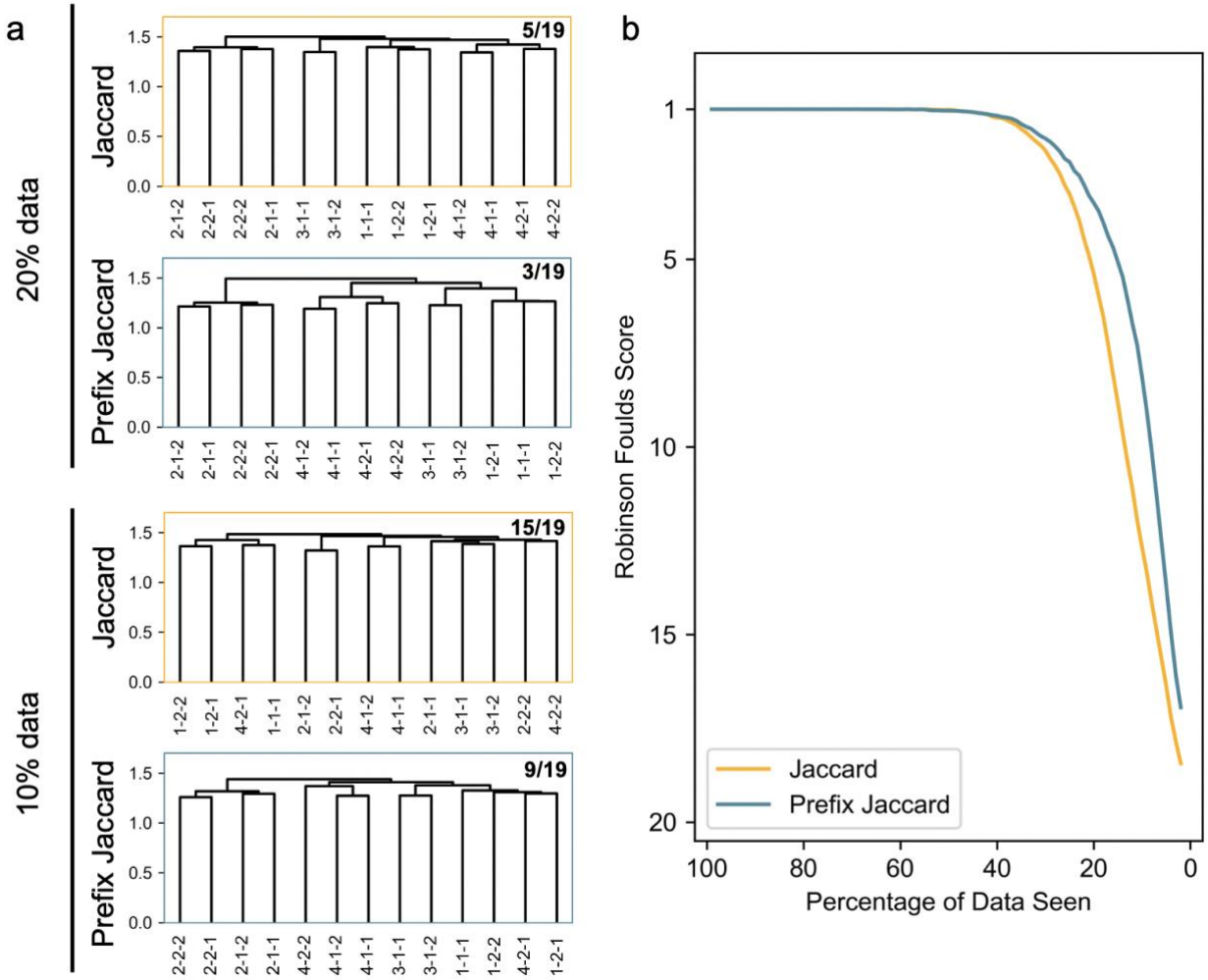
Supplementary Figure 4.2. Identification of efficient propagating pegRNAs. **a,** Heatmap of editing efficiencies for the top ten A→B propagator sequences identified in Supplementary Figure 4.1e. For each A→B propagator sequence (B_1-B_{10}), a library of signatures was tested, using PBS lengths ranging from 9-15 nts (row labels). **b,** Heatmap of editing efficiencies for pegRNAs that target the top ten propagator sequences shown in (**a**) and insert the 20-bp 6xHis sequence (B→A). Libraries encoding pegRNAs with all possible signatures, at each PBS length from 9-15 nts (row labels), were transfected into $293T_{6xHis}$ cells in which the corresponding B sequence had been installed at the recording locus. **c,** A→B and B→A pegRNAs can achieve propagation. Plasmids expressing PE2 and the top-performing pegRNA libraries identified in (**a**) and (**b**) were transfected into $293T_{6xHis}$ cells. The lengths of insertions at the recording locus were assayed after 2 rounds of transfection over 6 days. Insertions of 20 bp were considered to correspond to 1 round of recording, 40 bp to 2 rounds, and so forth.



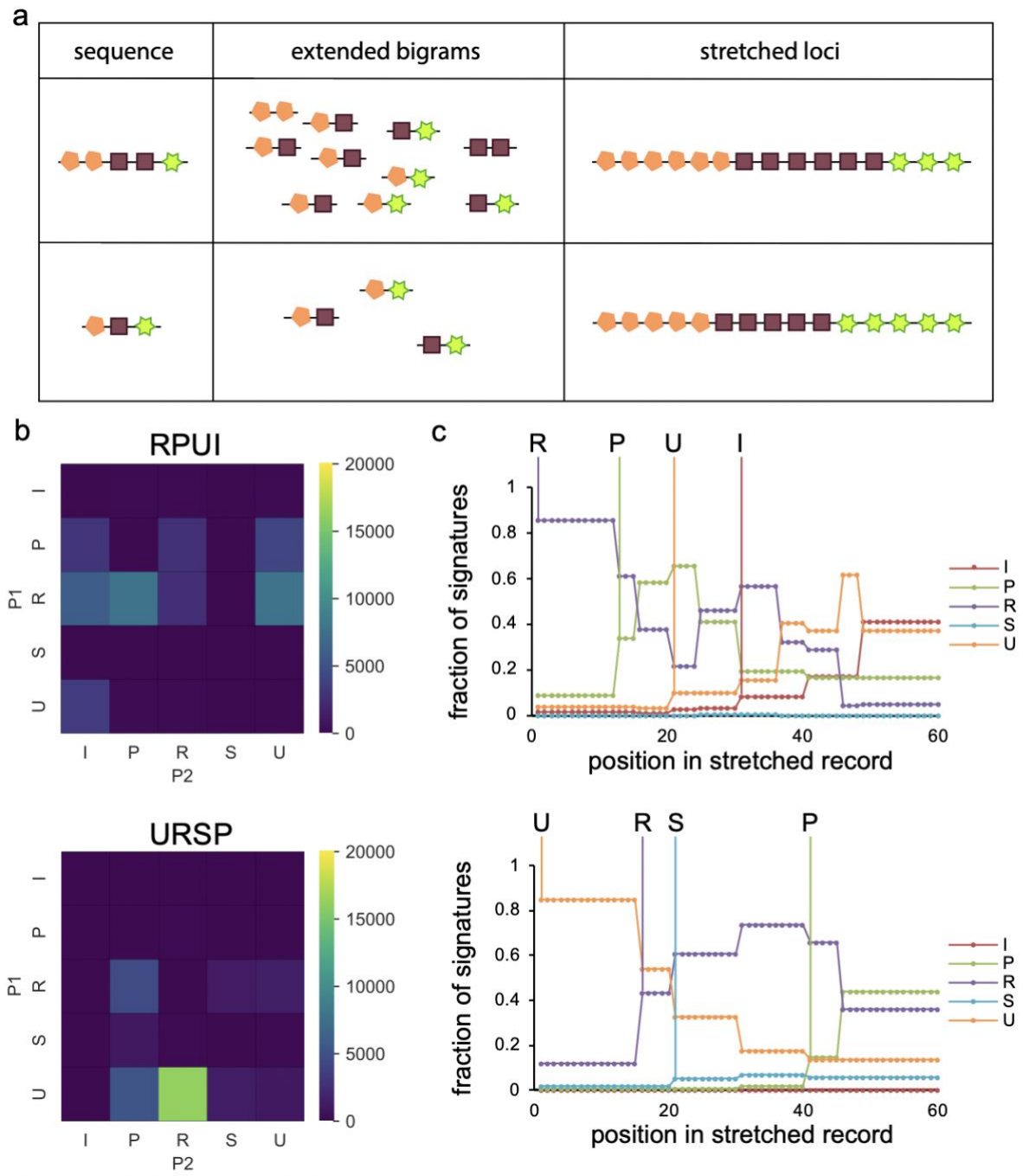
Supplementary Figure 4.3. Cannibalization by peCHYRON. **a,** Editing at the recording locus and pegRNA genes in peCHYRON cells. Prime editor and pegRNAs were integrated into the genome of $293T_{6xHis}$ cells as PiggyBac Transposase cargo, and integrants were selected with puromycin. 45 days after transfection, the recording locus (where editing is desired) and all pegRNA-expressing loci (where cannibalization occurs) were sequenced. The graph indicates the proportion of loci that were edited. Data from loci expressing A→B or B→A pegRNAs from each of three types of U6 promoters, human (h), mouse (m), and bovine (b), are shown. **b,** Cartoon showing cannibalization of B→A pegRNA gene by the A→B pegRNA. After transcription and coupling with prime editor, the spacer sequence of the A→B pegRNA can recognize either the target site at the recording locus, or the PAM-adjacent target site on the antisense strand of the B→A pegRNA gene. Note the B→A pegRNA only inserts 17 nt of the target sequence recognized by the A→B pegRNA, but the complete 20-bp target site and PAM are present in the B→A pegRNA gene because prime editing requires a scaffold sequence that complements with the region downstream of the nick, to enable flap equilibration after reverse transcription. This cartoon only shows cannibalization of the B→A pegRNA gene, but cannibalization can also occur in the inverse direction (the A→B pegRNA gene is cannibalized by the B→A pegRNA).



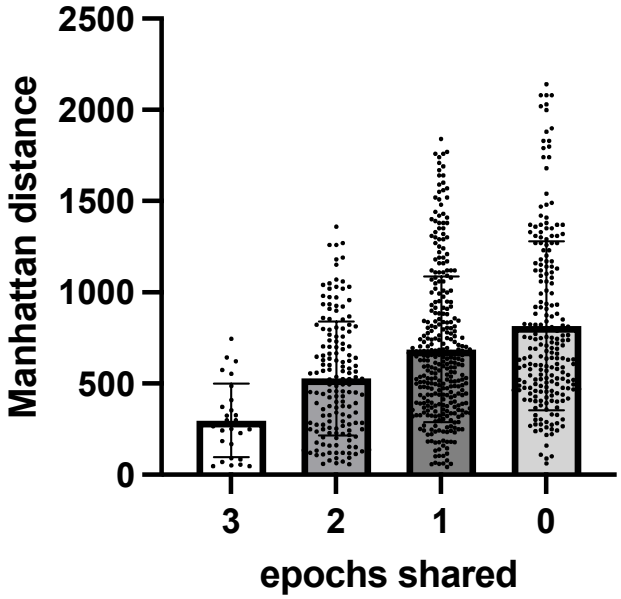
Supplementary Figure 4.4. Propagation of peCHYRON recording. **a,** Detailed breakdown of cannibalization data from Figure 4.3a, relevant to the experiment described in Figure 4.3a-c. The nature of cannibalization events– whether they produced a longer or shorter version of the pegRNA genes– is indicated.



Supplementary Figure 4.5. Comparison of lineage reconstruction methods. **a,** Reconstruction of culture splitting patterns from a cell lineage tracing experiment with random downsampling was more accurate when our prefix Jaccard similarity index was used. For each reconstruction, we calculated the Robinson Foulds score, which is the number of changes required to transform the reconstruction into the ground truth tree. The score for each reconstruction is shown in the upper right, divided by the highest possible score (19). **b,** Average Robinson Foulds scores for reconstructions using our prefix algorithm (blue) and the Jaccard distance (orange). This was calculated by downsampling the data to a certain percentage (x-axis) and performing a reconstruction which was then compared to the true tree. Each point was obtained by averaging over 1000 repeats. Note, perfect reconstruction in our case returns a score of 1 because we compare a rooted constructed tree to an unrooted ground truth tree.



Supplementary Figure 4.6. peCHYRON enables temporally-resolved event recording. **a,** Methods of peCHYRON sequence analysis. Any record can be deciphered by considering the extended bigrams of signature mutations it contains, or by stretching the record to a constant length, so it may be pooled with all other records from a population of cells and used to calculate the proportion of each signature at each position along the stretched records. **b,** Bigrams of recorded signatures reflect the order of transfections. $293T_{6xHis}$ cells, in which PEmax and the protective nicking sgRNA were stably expressed, were transfected with pairs of pegRNAs during 4 epochs of 3-6 days each. After the 4^{th} epoch, recording loci were sequenced. Heatmaps show the relative abundance of all possible ordered pairs; 3-letter signatures are represented by a single letter. The true epoch pattern is above each heat map. P1, first signature in a bigram. P2, second signature in a bigram. **c,** For the experiment shown in (b), results were analyzed and plotted as in Figure 5c, using computational stretching.



Supplementary Figure 4.7. Relationship between similar peCHYRON records and similar cell history. For the experiment shown in Figure 5c, all samples in biological replicate 1 were compared to all samples in biological replicate 2 by computing the pairwise Manhattan distances between the compositions of stretched records for each sample. For each pair of samples, the number of shared ordered epochs was also calculated. For example, the samples DID and DIC share two epochs, while the samples DID and IDI share 0 epochs. Each point represents a single pairwise Manhattan distance, bars represent the mean, and error bars represent the standard deviation. Results are significant by one-way Brown-Forsythe ANOVA test ($p<0.0001$, $F=32.94$, $DFn=3$, $DFd=559.4$), which was used because 3 of the groups failed the Shapiro-Wilk test of normality.