Ordered insertional mutagenesis at a single genomic site enables lineage tracing and analog recording in mammalian cells

Theresa B. Loveless*[1,2], Joseph H. Grotts*[1], Mason W. Schechter[1], Elmira Forouzmand[3], Bijan S. Agahi, Courtney K. Carlson[1], Guohao Liang[1], Xiaohui Xie[3], Chang C. Liu[1,4,5,6 #]

[1]Department of Biomedical Engineering, University of California, Irvine, Irvine, CA 92697, USA
[2]NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine, Irvine, CA 92697, USA
[3]Department of Computer Science, University of California, Irvine, Irvine, CA 92697, USA
[4]Department of Chemistry, University of California, Irvine, Irvine, CA 92697, USA
[5]Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine, CA 92697, USA
[6]Center for Complex Biological Systems, University of California, Irvine, Irvine, CA 92697, USA

* These authors contributed equally to this work.
[#]Correspondence to ccl@uci.edu

Summary

The study of intricate cellular and developmental processes in the context of complex multicellular organisms is difficult because it may require the non-destructive observation of thousands, millions, or even billions of cells deep within an animal. To overcome this difficulty, several groups have recently reported CRISPR-based DNA recorders that convert transient cellular experiences or processes into durable genomic mutations, which can then be read by next-generation sequencing in high-throughput. However, existing DNA recorders rely primarily on the accumulation of CRISPR-induced deletion mutations, which can be problematic because in the limit of progressive deletion mutations, no record remains. Here, we present a high-information DNA recorder that accumulates insertion mutations in temporal order at a single locus. Our recorder, called CHYRON (Cell HistorY Recording by Ordered iNsertion), can be applied as an evolving lineage tracer as well as a cellular stimulus recorder. As a lineage tracer, CHYRON allowed us to perfectly reconstruct the lineage relationships among 16 populations of human cells descended from four starting populations that were subject to a series of splitting steps. In this experiment, CHYRON progressively accumulated and retained insertions in 20% percent of cells such that the average length of insertion generated was 8.4 bp (~15 bits), reflecting high information content. As a stimulus recorder, we show that when the CHYRON machinery is placed under the control of a stress-responsive promoter, the frequency and lengths of insertions reflect the dose and duration of the stress. With further engineering of CHYRON's components to increase encoding capabilities and reduce loss, CHYRON's special ability to progressively accumulate insertion mutations should lead to single-cell-resolution recording of lineage and other information through long periods of time in complex animals or tumors, ultimately providing a full picture of mammalian development.

Keywords

DNA recording, lineage tracing, cellular barcoding, terminal deoxynucleotidyl transferase (TdT), hgRNA, stgRNA, CRISPR, Cas9

Introduction

Non-destructive observation of living organisms is a cornerstone of biology. Over time, our ability to observe ever-smaller organisms, individual cells within multicellular organisms, and molecules within cells has improved progressively with advances in microscopy and the continuing development of genetically-encoded labels that can be imaged non-destructively (*e.g.*, GFP). However, live imaging of single cells in intact organisms is still severely constrained by context and scale. Animals, for example, tend to be opaque and even when developmental processes are accessible to microscopy (McDole et al., 2018), cell tracking poses significant computational and data management challenges when the number of cells reaches only tens of thousands (McDole et al., 2018). An alternative paradigm to non-destructive direct observation is DNA recording. In DNA recording, transient cellular events are engineered to trigger permanent mutations in a cell's own genome (**Figure 1A-B**). Since DNA is both durable and propagating, and since the throughput of DNA sequencing is in the hundreds of millions of unique DNA molecules, the long-term behavior of cells could be stored as mutations in DNA and read out later at unprecedented depth. Although the reading step is destructive, recording is not, creating an effective alternative to direct observation that can scale to millions of cells in opaque model animals such as mice.

The paradigm of DNA recording has recently seen a transformation with the development of genetically-encoded CRISPR-based systems that drive rapid mutational accumulation at neutral loci in a cell's genome (Chan et al., 2019; Frieda et al., 2017; Hwang et al., 2019; Kalhor et al., 2018; Kalhor, Mali, & Church, 2017; McKenna et al., 2016; Perli, Cui, & Lu, 2016; Raj et al., 2018; Schmidt, Zimmerman, Wang, Kim, & Quake, 2017; Sheth & Wang, 2018; Sheth, Yim, Wu, & Wang, 2017; Shipman, Nivala, Macklis, & Church, 2016; 2017; Spanjaard et al., 2018; Tang & Liu, 2018). When the activity of such systems is linked to the presence of an arbitrary biological stimulus, accumulated mutations become a record of the strength and duration of exposure to the stimulus (Frieda et al., 2017; Perli et al., 2016; Sheth et al.,

2017); and when activity is constitutive, accumulated mutations capture lineage relationships among individual cells (Chan et al., 2019; Frieda et al., 2017; Kalhor et al., 2017; 2018; McKenna et al., 2016; Raj et al., 2018; Schmidt et al., 2017; Spanjaard et al., 2018). Two recent architectures for CRISPR-based recording systems are particularly amenable to recording in extremely large numbers of mammalian cells. The first architecture relies on arrays of *Streptococcus pyogenes* Cas9-target sites (Chan et al., 2019; Junker et al., 2016; McKenna et al., 2016; Schmidt et al., 2017). Here, Cas9 targets random elements of the array to generate insertions or deletions (indels) at array elements. The progressive accumulation of indels across the array yields a mutant array in each cell that can be used to infer lineage relationships or a cell's history of exposure to a stimulus. The second architecture relies on a self-targeting (Perli et al., 2016) or homing guide RNA (Kalhor et al., 2017) (hgRNA) that directs Cas9 to the very locus from which the hgRNA is expressed (**Figure 1C**). Here, the hgRNA locus changes over time to generate a pattern of indels at the locus that reflects lineage information or exposure to stimuli. These two types of systems have been used to identify the early and late embryonic origin of thousands of cell lineages in adult zebrafish (McKenna et al., 2016; Raj et al., 2018; Spanjaard et al., 2018), to study embryogenesis in mouse (Chan et al., 2019), and to record inflammation exposure in 293T cells implanted into mice treated with lipopolysaccharide (Perli et al., 2016). However, these systems suffer from a common and critical problem: the primary repair outcome of CRISPR nuclease-induced double-strand breaks (DSBs) in mammalian cells are deletions (T. Wang, Wei, Sabatini, & Lander, 2014). Since the progressive accumulation of deletions at a single locus will quickly corrupt or remove previous deletions (Kalhor et al., 2017; Perli et al., 2016), these DNA recording systems are limited in their information capacity and durability.

An ideal DNA recorder should be able to undergo continuous mutation without loss of information over time. The conceptually simplest way to achieve this is by making a DNA recorder that only accumulates insertion mutations in succession. We present a major step toward such a recorder by constructing a mutating DNA barcode called CHYRON (Cell HistorY Recording by Ordered iNsertion) (**Figure 1A-B**). CHYRON combines a Cas9 nuclease with an hgRNA and a DNA writer, terminal deoxynucleotidyl transferase (TdT) (Landau, Schatz, Rosa, & Baltimore, 1987; Pryor et al., 2018), which we show can efficiently insert random nucleotides (nts) at Cas9-induced DSBs. These inserted nts are then incorporated into the repaired DSB to produce a durable insertion mutation consisting of random base pairs (bps). Since the hgRNA repeatedly directs Cas9 to cut its own locus at a defined location relative to the PAM (Kalhor et al., 2017; Perli et al., 2016), cycles of cutting, nt insertion by TdT, and repair cause continuous and ordered insertional mutagenesis (**Figure 1A-B**). We describe the successful implementation of CHYRON and apply it to lineage reconstruction and recording of hypoxia. We find that the information generated at a single <100 bp CHYRON locus is sufficient to reconstruct the relatedness of populations containing thousands of lineages as well as to report on the duration and dose of exposure to a hypoxia mimic. This opens up the possibility of following mammalian development with unprecedented ease and depth, with the ability to distinguish the hundreds of millions of cells produced during mammalian development or profiling the heterogeneous responses of each cell in a population to unevenly distributed or dynamic stresses.

Results and Discussion

TdT mediates random insertional mutagenesis at Cas9 cut sites

The core functionality of CHYRON relies on insertional mutagenesis by TdT. Therefore, we first tested whether a single round of Cas9 cutting and DSB repair could be intercepted by TdT to generate insertions. In ~10^5 HEK293T cells, we targeted Cas9 to a genomic locus with or without the ectopic expression of TdT. We then analyzed repair outcomes by PCR amplification of the target locus followed by next-generation sequencing (NGS), taking care to capture substitution, deletion, and insertion mutations equally (see **Methods**). We found that without TdT, the dominant mutation present at the Cas9 target site was a deletion (84%) or 1 bp insertion (13%) (**Figure 2A**), consistent with previous literature. However, with TdT, the dominant mutations were insertions (74%) (**Figure 2A**), with an average length of 2.8 bp (**Figure 2B**). This dramatic shift in repair outcomes in a single round of editing meant that once

combined with an hgRNA, CHYRON should be able to undergo insertional mutagenesis over multiple rounds with minimal sequence corruption or loss through deletions.

TdT-mediated insertions must have a wide diversity of possible sequences for CHYRON to be an effective recorder. We found that TdT-mediated insertional mutagenesis generated an average insertion length of 2.9 bp (5.2 bits) per round (**Table S1**). To calculate these values, we characterized the single-round length of insertions at Cas9 targets in the presence of TdT. As shown in **Figure 2B**, insertions were commonly 1-4 bp, with some longer insertions that raised the average length to 2.9 bp. To measure biases in the inserted bp in order to determine the average number of bits of information in each bp inserted, we designed single guide RNAs (sgRNAs) to recruit Cas9 to a panel of genomic sites. Our choice of target sites was made to explore all 16 possible pairs of the -4 and -3 nts relative to the PAM, since Cas9 canonically creates a blunt cut between these -4 and -3 positions and there is a known influence of these nts on editing outcomes. We found that single-bp insertion outcomes were likeliest to have the identity of the -4 nt, a bias that was independent of TdT (**Table S1**) and that could be explained by staggered Cas9 cuts (-4 relative to the PAM in the non-template strand and +3 relative to PAM in the template strand) (Gisler et al., 2019; Jinek et al., 2012; Zuo & Liu, 2016) or by the requirement for cohesive ends in order to promote re-ligation of a DSB. However, in longer insertions, we found that the identity bias of inserted bp was driven by TdT, which is known to prefer the insertion of Gs *in vitro* and during VDJ recombination (Motea & Berdis, 2010). This resulted in our overall observation that Gs and Cs comprised ~70% of nts inserted (**Figure 2C** and **Table S1**). Our results also indicated that TdT added to both 3' ends of the DSB, since we saw a similar preference for both Gs and Cs in longer insertions. However, some target sites resulted in insertions with a strong bias for G over C or vice versa, suggesting that TdT prefers to add nts to one DNA terminus at a DSB over the other in a sequence-dependent manner. To determine the information content of TdT-mediated insertions, we calculated the entropy per bp for each insertion length at all 16 genomic target sites we tested (**Table S1**). (This calculation implies that the biases of bps inserted were independent of each other, which is a reasonable approximation of what we observe.) From these entropies per bp, we calculated the average entropy per round of editing at each site by weighting by the frequency of each insertion size. Finally, we averaged this entropy per round calculation and the average length of insertions over all 16 sites to get an average entropy per round of 5.2 bits and an average insertion length of 2.9 bp. Therefore, we conclude that TdT-mediated insertions encode an average of 1.78 bits of information per bp (**Table S1**), compared to 2 bits if all 4 bases were added randomly.

CHYRON$_{20}$ accumulates ordered insertion mutations in multiple rounds

A recording locus should autonomously accumulate mutations over multiple rounds of activity. To achieve continuous rounds of insertional mutagenesis, we combined TdT with an hgRNA locus (Perli et al., 2016) (Kalhor et al., 2017) to establish CHYRON. Because Cas9-induced DSBs are consistently generated between the -4 and -3 nts relative to the PAM of the hgRNA locus (**Figure 2D**), rounds of TdT-mediated insertion mutations should follow in order when repeated. This makes CHYRON an ideal recording locus because 1) new insertions will neither remove nor corrupt previous insertions and 2) insertions are directionally arranged in the exact order in which they are added, simplifying inference of lineage or stimuli exposure from the mutational information recorded.

To demonstrate repeated and ordered insertional mutagenesis, we integrated an hgRNA locus, including a 20 nt spacer, at a single site in 293T cells. When this cell line containing CHYRON$_{20}$ – we use the subscript to distinguish this specific instantiation of CHYRON that has a 20 nt hgRNA spacer from ones discussed later – was transfected with a plasmid expressing Cas9 and TdT for three days, insertions accumulated at the locus as expected (**Figure 3A**). (For simplicity, we will refer to hgRNA loci as CHYRON loci when in the presence of TdT, and to the cell line bearing the integrated CHYRON$_{20}$ locus as 293T-CHYRON$_{20}$.) As a comparison, we carried out a similar experiment where a genomic locus with the same spacer sequence as CHYRON$_{20}$ was targeted by an sgRNA, thereby allowing for only a single round of editing (shown in **Figure 2A-B**). We found that 1-2 bp insertions were less abundant and longer insertions were more abundant at the CHYRON$_{20}$ locus compared to the genomic locus targeted by an

sgRNA. This difference strongly suggested that the $CHYRON_{20}$ locus was edited in multiple successive rounds. In order to show multi-round editing conclusively, we isolated 293T-$CHYRON_{20}$ cells that had acquired an insertion at the $CHYRON_{20}$ locus to near-clonality, then transfected them with Cas9 and TdT again. Although further editing was inefficient, new insertions were abundantly observed and all new insertions were found precisely downstream of the original insertion (**Figure 3D-E** and **Figure S3B-C**).

$CHYRON_{20}$ gave our basic desired CHYRON behavior, progressively generating 8.1 bits of information on average via ordered insertions of short random bp stretches, but it was clear $CHYRON_{20}$ could be improved. For example, we deduced that $CHYRON_{20}$ only underwent approximately two rounds of editing, because the proportion of deletion-containing sequences was ~35% of the converted $CHYRON_{20}$ loci and we knew that a single round of editing generates ~26% deletions (**Figure 2A**), requiring two rounds to give ~35%. We also found that when 293T-$CHYRON_{20}$ cells were transfected with Cas9 and TdT over 9 days, the average insertion length plateaued at 4.9 bp, which was already reached after 6 days of Cas9/TdT expression (**Figure 3B**). This was shorter than the average 5.8 bp length that we expected for two rounds, with the discrepancy suggesting that shorter initial insertions were disproportionately likely to continue to edit. Therefore, we sought to improve $CHYRON_{20}$ to be capable of more rounds of activity, which should result in greater potential diversity of insertion sequences.

$CHYRON_{16i}$ accumulates an average of 8.4 inserted bps over an average of three rounds

The failure of insertions in 293T-$CHYRON_{20}$ cells to extend further than an average of 4.9 bp has two likely explanations: silencing of the CHYRON locus or reduced efficiency of the hgRNA through increased length and/or secondary structure associated with continued rounds of TdT-mediated insertions. To address these potential problems, we created two new cell lines, 293T-$CHYRON_{20i}$ and 293T-$CHYRON_{16i}$, both of which have the CHYRON locus flanked by chromatin insulator sequences (M. Liu et al., 2015) and integrated at the *AAVS1* safe harbor locus in 293T cells. $CHYRON_{20i}$ starts with a 20 nt spacer, which operates at peak activity, while $CHYRON_{16i}$ starts with a 16 nt spacer, which is predicted to have very low activity initially (Fu, Sander, Reyon, Cascio, & Joung, 2014), but has more room to accumulate insertions before it reaches lengths that prohibitively reduce editing efficiency (Kalhor et al., 2017; Perli et al., 2016). These new CHYRONs significantly outperformed the original $CHYRON_{20}$. Specifically, 293T-$CHYRON_{20i}$ cells, in contrast to 293T-$CHYRON_{20}$ cells, continued to accumulate longer insertions throughout the entire 9-day time course and resulted in $CHYRON_{20i}$ loci that reached a final length of 5.7 generated bps on average (**Figures 4A** and **4B**); and 293T-$CHYRON_{16i}$ cells, even though they had a lower overall editing efficiency due to the low starting activity of the shorter sgRNAs, continued to accumulate insertions to an average length of 8.4 generated bps, encoding 15.3 bits of information (**Figures 4A** and **4B**). In short, the $CHYRON_{16i}$ locus progressively generated far more information at a single site than any previous DNA recorder with ~15 bits, or ~33,000 possibilities, and should be broadly useful.

$CHYRON_{16i}$ allows the reconstruction of relationships among 16 populations containing thousands of lineages

To mimic a process of growth and differentiation over several days in a setting in which we could know the ground truth of lineage relationships among the cells, we 1) split 293T-$CHYRON_{16i}$ cells bearing the root $CHYRON_{16i}$ sequence into 4 wells, 2) expressed Cas9 and TdT for three days (approximately three doublings) to allow the cells to acquire insertions, 3) split each well into two, and 4) repeated steps 2 and 3 again to yield 16 final wells (**Figure 5A**). Cells in these final wells were allowed to grow for three days. The approximately three doublings between splits ensured that enough cells could acquire an insertion and then divide between splits, and the grow-out at the end of the experiment ensured that our recovery of unique sequences was high.

By subjecting cells in only the final wells to NGS of the CHYRON locus, we were able to robustly generate a perfect reconstruction of the full splitting procedure. This was done using the presence of shared sequences between pairs of wells to calculate relatedness (similarity) and a standard agglomerative

hierarchical clustering method to generate the tree from pairwise similarities (**Figure 5B**). We also developed a novel reconstruction algorithm, the "average prefix method," to take advantage of the ordered nature of insertions at the CHYRON locus. The average prefix method also accurately reconstructed the splitting procedure (**Figure 5C** and **Figure S5D**). The ease and accuracy of lineage reconstruction in this rather complex experiment – one population expanded to sixteen, and 10^4 cells expanded to ~4 x 10^7 (**Figure 5A**) – predicts that CHYRON should be a powerful system for deep lineage reconstruction of developmental processes involving substantial proliferation and fate changes.

It is instructive to note that perfect reconstruction resulted from the analysis of CHYRON insertions at least 8 bp in length (**Figure 5B**) and the quality of the reconstruction decreased slightly when including shorter insertion sequences (*e.g.* 7 bp, **Figure S5C**). In addition, the exclusion of shorter sequences was essential for accurate reconstruction when the abundance cutoff was set low (**Figure S5D**) or when the data were downsampled (**Figure S5E**). Why did the inclusion of shorter insertion sequences reduce reconstruction robustness? The answer to this question results from the interplay between convergence and sampling that should apply to all DNA-recording-based lineage reconstruction studies.

Let us consider a well (Well A), its closest relative (Well B), and a totally unrelated well (Well X). Suppose we detect a specific CHYRON sequence (or a shared substring of the CHYRON sequence given the ordered nature of insertions in CHYRON) in only two of these three wells. Must the two wells be Well A and Well B? If the sequence is long, then it almost certainly must be, since a long CHYRON sequence will not arise independently by chance. However, if the sequence is short, then it is possible that the two wells sharing the sequence are, for example, Well A and Well X, and that the sequence independently generated in both wells. The critical issue, however, is that the sequence could be absent in Well B, because 1) the sequence was generated after Well A and Well B split from their common ancestor or 2) the sequence was present also in Well B but was not detected due to sampling inefficiencies. Therefore, the short sequence will preferentially assign relatedness to the unrelated wells over the related wells, reducing the accuracy of reconstruction. To minimize this effect, one must use sufficiently long CHYRON sequences and ensure sufficient sampling of the sequences in all wells, a conclusion generalizable to all lineage reconstruction studies from self-mutating DNA recording systems. (See **Supplemental Results and Discussion**, **Supplemental Figures S4** and **S5**, and **Supplemental Tables 2-4** for further analysis and testing of the effects of convergence and sampling on reconstruction quality, and the implications of this analysis on future work for CHYRON and the design of lineage tracing experiments from all DNA recorders.)

$CHYRON_{20}$ and $CHYRON_{16}$ can report the dose and duration of exposure to a hypoxia mimic

DNA recording systems have been used to log cellular exposure to biological stimuli by making mutation at the recording locus inducible by biological stimuli of interest. For example, mutational accumulation at hgRNAs have been linked to inflammation exposure (Perli et al., 2016). However, in such cases, recording has been digital from the perspective of a single cell: the information used from the hgRNA is whether it is mutated or not. The dose or duration of the stimulus being recorded is therefore reflected at the population level, in the proportion of the population that bears a mutated hgRNA. While CHYRON can do the same, CHYRON also offers the possibility of acting as a compact recorder that is analog from the perspective of a single cell. This is because the CHYRON locus progressively accumulates insertions and rarely obtains deletions, so a CHYRON locus should grow monotonically longer as the cell is exposed to the stimulus for a longer period of time, or at a higher dose.

To test stimulus recording with CHYRON, we linked insertional mutagenesis at CHYRON to hypoxia, which triggers adaptive responses that affect a wide range of cellular behaviors including ones important for tumor evolution and metastasis (Rankin & Giaccia, 2016; Semenza, 2012). We created a construct in which the expression of Cas9 and TdT is under the control of the 4xHRE-YB-TATA (Ede, Chen, Lin, & Chen, 2016) promoter, and in which Cas9 is additionally fused to an oxygen-dependent degron domain. We transfected 293T-$CHYRON_{20}$ and 293T-$CHYRON_{16}$ (which bears a 16-nt-spacer hgRNA integrated at the *AAVS1* locus, as in $CHYRON_{16i}$ but without insulators) with this construct, and then exposed them

to three different concentrations of the hypoxia mimic DMOG for five different durations. In both cell lines, the proportion of the population bearing insertions increased with dose and duration of DMOG treatment (**Figures 6A** and **6B**). In the case of $CHYRON_{16}$, the average length of insertions also increased with duration of DMOG exposure for the sequences that were mutated (**Figure 6C**). In other words, CHYRON is capable of recording exposure to stimuli in a manner that is digital (**Figures 6A** and **6B**) or analog (**Figure 6C**), where the latter of these modes can in principle provide information on the experience of each single cell. We note that currently, the dynamic range of analog recording achieved with CHYRON is narrow (**Figure 6C**), but with further development, we expect CHYRON will be an ideal system for capturing detailed cellular histories at single-cell resolution.

Current and future capabilities of CHYRON

There are three unique features of the CHYRON architecture that we believe will lead to its broad application and motivate its continued development in our and other labs. First is the high information content and density of CHYRON. CHYRON is able to diversify a very compact recording locus, consisting of a single site that is repeatedly modified, so that the locus can bear a unique sequence in each of tens of thousands of cells. This capability may be especially important for applications where it is difficult to capture all cells that might be related to each other, in which case a DNA recorder with a high information content is necessary to limit the possibility of misleading convergent sequences in unrelated cells. Second is the property that CHYRON records information by generating an ordered accumulation of random insertions. Unlike deletions and substitutions, pure ordered insertions gain information without corrupting or removing previous information, which is ideal for a DNA recorder. The ordered nature of insertions generated also introduces the possibility that, if TdT can be engineered to add different types and lengths of nts deterministically and the activity of the different TdTs can be coupled to different cellular stresses or the cell cycle, the CHYRON locus would record the relative timings of the different stresses in the cell's history or even provide an accurate count of cell divisions. The latter may enable single-cell-resolution lineage reconstruction from sparsely sampled CHYRON sequences, a goal we are actively pursuing. Third, CHYRON is a high-information DNA recorder that uses only a single genomic site. Because the recording site and Cas9/TdT machinery can be encoded at a single locus, CHYRON can be readily transplanted into new cell or animal lines.

In its current form, CHYRON is likely the best option for reconstruction of the relationships between thousands of cell lineages resulting in millions of cells. However, additional work is necessary to reach its full potential. Currently, TdT is recruited to the Cas9 cut site through its natural interaction with the DSB repair machinery. As a result, it is likely recruited to all DSBs in the cell and will act as a mutagen over long periods of expression. To ensure that normal development is not perturbed by this increased mutagenesis, TdT should be engineered so that it is specifically recruited only to the CHYRON locus. As shown in this study, we could not observe increased activity of TdT *in cis* when it is fused to Cas9 (**Figure S2C**). Mutations of TdT that prevent its binding to the DNA repair machinery have been reported (Mahajan et al., 1999), and a better strategy may be to fuse these mutants to a protein that remains at the DSB after Cas9 dissociates. The information-encoding capacity of CHYRON, although unprecedentedly high for a single site, is limited by the declining efficiency of the hgRNA as it grows longer. The reduced efficiency likely arises from a combination of guide RNA length and secondary structure in the critical seed region. Several approaches could address these issues, namely engineering Cas9 to better-tolerate these types of sequences or switching to use a different nuclease that cuts further from its seed region. Finally, the ~25% rate of deletion per Cas9 cut will still eventually lead to information loss and inactivation of all CHYRON locus in the limit of truly continuous recording. However, recruitment of factors that manipulate the balance of DSB repair pathways at the Cas9 cut site could reduce deletions significantly. The future development of CHYRON will be enhanced by the wide interest in engineering new capabilities into its protein components – a CRISPR nuclease (Sheth & Wang, 2018) and TdT, in which there has been considerable recent interest as a tool for *in vitro* DNA synthesis (Lee, Kalhor, Goela, Bolot, & Church, 2018; Palluk et al., 2018). Techniques that use polymerases (Glaser et al., 2013; Zamft et al., 2012), including TdT (Bhan et al., 2019), to record time-series information on DNA synthesis

timescales *in vitro* could also be merged with CHYRON. In short, we predict that the unique components of CHYRON and the promise of the CHYRON architecture in reaching fully continuous recording of biological stimuli or lineage relationships at single-cell resolution *in vivo* will spur its continued development and application.

Materials and Methods

Plasmid cloning. Cloning was done through standard Gibson assembly and Polymerase Chain Reactions (PCRs) were performed with Q5 Hot Start High-Fidelity DNA Polymerase or Phusion Hot Start Flex DNA Polymerase (New England BioLabs). All primers were purchased from Integrated DNA Technologies (IDT) and PCR reagents were provided by NEB. All transformations were done in Top10 *E. coli* (ThermoFisher Scientific), unless otherwise stated.

The human codon-optimized *S. pyogenes* Cas9 DNA sequence was PCR amplified from hCas9, which was a gift from George Church (Addgene plasmid # 41815 (Mali et al., 2013)). Either an XTEN linker or a T2A self-cleaving sequence along with TdT were cloned onto the C-terminus of Cas9. The XTEN linker was cloned through PCR from pCMV-BE3, which was a gift from David Liu (Addgene plasmid # 73021 (Komor, Kim, Packer, Zuris, & Liu, 2016)). The T2A self-cleaving sequence was inserted through PCR by designing primers with overhangs containing the T2A sequence. The TdT DNA fragment was amplified from the cDNA of an acute lymphoblastic leukemia cell line through PCR and this entire insert was cloned into a pcDNA3.1 backbone, yielding Cas9-XTEN-TdT or Cas9-T2A-TdT. An additional construct was cloned through PCR and Gibson assembly that fused TdT to the N-terminus of Cas9 through the XTEN linker, yielding TdT-XTEN-Cas9-XTEN-TdT. Cas9-containing constructs with the pcDNA3.1 backbone were transformed into XL10-Gold Ultracompetent *E. coli* (Agilent #200315).

Additional Cas9-TdT constructs containing different linkers were cloned through restriction enzyme digestion and ligation. All restriction enzymes and T4 DNA Ligase were purchased from NEB. All vector digestions were treated with alkaline phosphatase, calf intestinal (CIP) from NEB after complete digestion by the restriction enzymes. A base construct was cloned first through Gibson Assembly to add restriction sites to the original Cas9-XTEN-TdT construct: NheI-SfiI-Cas9-KpnI-XTEN-SexAI-TdT. The base construct was digested with KpnI and SexAI. Three separate PCR's were performed to yield a 5xFlag or 5xGSA linker product with KpnI and SexAI restriction sites. The 5xFlag was present on a gBlock (IDT) and the 5xGSA linker (4 repeats of the sequence GSAGSAAGSGEF and a final repeat with the sequence GSAGSAAGASGEGRP (Waldo, Standish, Berendzen, & Terwilliger, 1999)) was ordered on a minigene (IDT). These two inserts were digested with the appropriate enzymes and ligated individually into the KpnI and SexAI digested base construct yielding Cas9-5xFlag-TdT or Cas9-5xGSA-TdT.

Additional PCR on the 5xFlag gBlock yielded 5xFlag flanked by KpnI sites, which was digested and then ligated into Cas9-5xFlag-TdT, resulting in Cas9-10xFlag-TdT. The same was performed on the minigene to amplify the GSA linker and ligated into Cas9-5xGSA-TdT, to yield Cas9-10xGSA-TdT. To clone Cas9-15xFlag-TdT, a subsequent PCR was performed on the gBlock to amplify the 5xFlag sequence with FseI restriction sites. The FseI restriction-enzyme recognition sites were used to ligate the 5xFlag sequence into Cas9-10xFlag-TdT, yielding Cas9-15xFlag-TdT.
Hypoxia inducible constructs were cloned through adding a 4x hypoxia-response element (HRE)-YBTATA promoter to drive the expression of Cas9-T2A-TdT. The 4xHRE sequence was PCR amplified from 4xHRE_v2_YB-TATA-Gluc-CMV_dsRed (Ede et al., 2016), which was a gift from Yvonne Chen. In addition, the oxygen-dependent degron (ODD) was cloned onto the C-terminus of Cas9. The ODD sequence was amplified from HA-HIF1alpha-wt-pBabe-puro, which was a gift from William Kaelin (Addgene plasmid # 19365 (Yan, Bartz, Mao, Li, & Kaelin, 2007)). This plasmid also includes blasticidin resistance (not used in this work), which was cloned from pLenti CMV Blast DEST (706-1) backbone, which was a gift from Eric Campeau and Paul Kaufman (Addgene plasmid # 17451 (Campeau et al., 2009)).

Control plasmids were cloned containing catalytically-dead TdT (dTdT). The dTdT DNA fragment was prepared through introduction of the D343E and D345E mutations (Repasky, Corbett, Boboila, & Schatz, 2004; Yang, Gathy, & Coleman, 1994) into the wild-type TdT sequence. Two control plasmids were cloned through Gibson Assembly into the pcDNA3.1 backbone: Cas9-XTEN-dTdT and dTdT alone.

To clone single-guide RNA (sgRNA) plasmids, the spacer region of the desired sgRNA was inserted into the pSQT1313 expression plasmid, under the control of the U6 promoter, which was a gift from Keith Joung (Addgene plasmid # 53370) (Tsai et al., 2014). 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 *E. coli* (Lucigen) to allow for recombination to yield the desired variant sgRNA plasmid.

The homing gRNA (hgRNA) constructs contained the HEK293 site 3 sgRNA cassette with a GGG (instead of GTT) at the 3'-end of the spacer region and the complementary mutations in the opposite site of the hairpin (Kalhor et al., 2017; Perli et al., 2016). 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 site3 locus was cloned upstream and downstream of the hgRNA and selection marker region of the plasmid. The sequences of the flanks were PCR amplified 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 to be stably integrated into HEK293T cells upon transfection.

The promoter, expressed sequences, terminator, and homologous regions of each plasmid were verified by Sanger sequencing (Genewiz), and all plasmids to be used for transfection were purified with HP GenElute Midi or Mini kits (Sigma # NA0200 and NA0150).

Cell culture and transfection. All cell culture experiments were performed in HEK293T cells obtained from ATCC (CRL-3216). Cells were cultured in DMEM, high glucose, GlutaMAX^{TM} Supplement (Gibco #10566024), supplemented with 10% FBS (Sigma #12306C), at 37°C and 5% CO_2.

Transient transfections were performed by mixing DNA with Fugene (Promega #E2311) in serum-free DMEM, at a ratio of 1 μg DNA to 3 μL Fugene.

To create the 293T-CHYRON cell lines, the plasmid to integrate the hgRNA into 293T cells was digested with EcoRI, then purified on a silica column (Epoch #3010). 350 ng of this plasmid was mixed with 100 ng of MSP680, a plasmid expressing Cas9^{EQR}, a gift from Keith Joung (Addgene #65772 (Kleinstiver et al., 2015)), 50 ng of a plasmid expressing an sgRNA to a HEK293 site 3 or *AAVS1* site that can be cut by Cas9^{EQR}, and 1.5 uL Fugene. 293T cell were transfected in a 24-well dish, transformants were selected with puromycin (Invivogen #ant-pr-1), then a single colony was isolated in two rounds of dilution and colony picking.

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

Examining the insertion bias of TdT at varying cut sites. To test the insertion characteristics of TdT, 16 targetable sites were chosen that contained all combinations of each nucleotide at the -4/-3 position relative to the PAM (Supplementary Table 1). A six-well well of HEK293T cells were transfected with 0.4 μg of the specific sgRNA and either 0.92 μg of Cas9, 1.09 μg of Cas9-T2A-TdT, or 1.1 μg of Cas9-5XFlag-TdT. To normalize the total number of nucleotides transfected, the Cas9 and Cas9-T2A-TdT transfections were supplemented with 0.12 μg and 0.01 μg of pcDNA3.1-sfGFP, respectively. The

pcDNA3.1-sfGFP construct was cloned through Gibson Assembly of the superfolder GFP gene into pcDNA3.1. The cells were collected three days post-transfection and processed for DNA and protein as detailed below.

Long-term editing on the CHYRON locus. 293T-CHYRON$_{20}$, $_{20i}$, or $_{16i}$ cells were transfected in 6-well dishes with 2 μg of the Cas9-T2A-TdT construct. The editing took place for 1, 2, 3, 6, or 9 days after transfection. For the 6 and 9-day time point, 10% of the cells from the previous time point were used to seed a new culture to be transfected again with the same amount of DNA as the previous transfection. As a control, the same experiment was performed with a Cas9-T2A-TdT plasmid with a stop codon five amino acids into the TdT sequence. All time points were collected as a single well of a confluent six-well plate (Falcon # 08-772-1B).

Two-step editing with Cas9 and TdT via isolation of single colonies. 293T-CHYRON$_{20}$ cells were transfected with equal amounts of plasmids expressing Cas9-5xFlag-TdT and free TdT, then diluted and single colonies picked. The CHYRON locus of these colonies was sequenced by the Sanger method, and six cell lines were chosen for further study. These six cell lines were grown in two 6-wells each. For each, one well was transfected with a plasmid expressing Cas9-T2A-TdT and the other well was untransfected. Three cell lines representing two insertions were found to be nearly clonal and successfully sequenced. All samples were collected and the CHYRON locus sequenced via UMI incorporation and NGS.

Lineage reconstruction assay and analysis. 5,000 293T-CHYRON$_{16i}$ cells were plated in each of 4 wells of a 384-well plate, then transfected the next day with a plasmid expressing Cas9 and TdT. Three days later, when they had expanded to approximately 86,000 cells per well, they were each split into two wells of a 24-well, allowed to attach for one day, then transfected again. Three days later, when each well had expanded to approximately 800,000 cells, each well was split into two wells of a 6-well dish. Three days later, all wells were collected and analyzed by amplicon sequencing without UMI incorporation.

For our initial analysis, we created a list of all insertion sequences in each well. Each insertion has an "abundance," based on the number of NGS reads that include that exact insertion sequence, and a length, equal to the number of bp added to the root sequence at the Cas9 cut site. We refined the list for each well to include only those insertions that met two criteria: (1) they were represented in at least 0.0139% of the non-deletion reads in the well and (2) the inserted sequence had a length of 8 to 15 bp. From this list of insertions, we created a binary vector for each well whose length was equal to the total number of insertion sequences with these criteria observed in any of the 16 wells in the experiment. The vector for each well contains a 1 for a particular insertion if that insertion was present in the refined list for that well, or a 0 if that insertion is absent. We used these vectors to calculate the Jaccard similarity between each pair of wells (Kalhor et al., 2017), then reconstructed the relationships using the UPGMA hierarchical clustering algorithm (https://github.com/scipy/scipy/blob/v1.2.1/scipy/cluster/hierarchy.py#L411-L490).

To determine the average prefix between two wells, we assess each insertion in the first well, pair it with its closest match in the second well, and record the length of the initial identical sequence (prefix) shared by those two insertions. After we perform this process for each insertion in both wells, we average the prefix lengths to determine the average prefix between those two wells. After calculating the average prefix shared by each pair of wells, we pair the two wells with the longest average prefix, then those with the next-longest average prefix, etc, until all sibling wells are paired. Then, the identical sequences from each pair of wells are collected the create a "pooled well," and those pooled wells are paired by the same process to reconstruct the relationships between cousins.

To visualize the relationships between wells as determined by the average prefix method, we converted the average prefix to a measure of distance by the following equation:
distance = 1/(average prefix – shortest prefix found in pairing)

Then, we arbitrarily set the shortest distance to 0.5 inches for visualization, and set the other distances proportionally.

All the analyses were done in python. The scripts are available on Gitlab at https://gitlab.com/__mason__/CHRYON.

Hypoxia recording assay. 293T-CHYRON$_{20}$ and 293T-CHYRON$_{16}$ cells were transfected in 6-well dishes with 2 μg of the 4XHRE-YBTATA-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 hours later. 14 hours 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.

Deep sequencing library preparation of a genomic locus (for Figures 2, S1, and S2). Genomic DNA was isolated with a QIAAmp DNA Mini Kit (Qiagen #51304) and the region targeted by Cas9 was amplified by PCR. The primers contained the Illumina adapters and a 5 – 7 nt sample-specific barcode (Supplementary Table 2). The PCR reaction 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. For the same genomic locus, each sample was normalized by signal intensity on a 0.9% gel and pooled into a single mixture, which was cleaned using a NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel #NC0389463). 10 ng per individual sample from the pooled clean product was sent to Quintara Biosciences or FornaxBio and ran 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).

Deep sequencing library preparation of the CHYRON locus (for Figures 5, S4, S5, and S6). Genomic DNA was extracted with the QIAmp DNA Micro Kit (Qiagen) and carrier RNA (for Figure S4) or the QIAmp DNA Mini kit and the entire recovery was used in the initial PCR. The initial PCR was performed for 25 cycles with Phusion HotStart Flex polymerase in GC buffer (New England Biolabs), each sample was purified with AMPure beads (0.9:1), then digested with PmlI (New England Biolabs) for 4 hours, then purified with AMPure beads again. Then the reamplification PCR was performed for 15-25 cycles in Q5 HotStart polymerase. The samples were pooled according to their estimated concentration on an agarose gel stained with ethidium bromide, then purified with AMPure beads, and cut with PmlI for an additional 4 hours. Finally, bands of the expected library size or up to 100 bp larger were gel-purified using a Macherey-Nagel PCR Clean-up kit. They were sequenced on an Illumina HiSeq 2500 using the PE100 kit at the UCI Genomics High Throughput Facility.

Unique molecular identifier incorporation for sequencing of the CHYRON locus (for Figures 3, 4, 6, and S3). To barcode individual cells containing the integrated CHYRON locus, unique molecular identifiers (UMIs) of 20 degenerate nucleotides were incorporated. Primers were ordered from IDT containing the following: the Illumina reverse adapter, a UMI, a 5 – 7 nt sample-specific barcode, and a stgRNA construct binding region (Supplementary Table 3). gDNA was isolated with the QIAmp DNA Mini Kit (Qiagen) and 600 ng of nucleic acid was used. The UMI incorporation reaction was run with Phusion Hot Start Flex DNA Polymerase (NEB) and under the following condition: 98°C, 5 min; (55°C, 30 s at a ramp rate of 4°C/s ramp rate; 72°C, 1.5 min) × 10. The reaction was enzymatically cleaned with Exonuclease I and Shrimp Alkaline Phosphatase (NEB) by incubating the sample and enzymes for 30 minutes at 37°C.

A downstream PCR was performed on the UMI incorporation step to amplify specific sequences that contained a UMI. The sample was run with a forward primer with the Illumina forward adapter, a 5 – 7 nt sample-specific barcode, and a stgRNA binding region, and a reverse primer complementary to the Illumina reverse adapter present on the UMI primer. The PCR was performed under the following conditions: 98°C, 3 min; (98°C, 1 min; 65°C, 30 s; 72°C, 30 s) × 35; 72°C, 1 min with a 2°C/s ramp rate. Products were purified on columns from the NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel) and individual samples were pooled based on equal molar ratios. Samples were further processed as for genomic sites for Figure 3A-C. For the rest of the experiments, samples were individually purified with AMPure beads (0.9:1), then reamplified for 15 cycles, pooled, and gel-purified including ~50 bp smaller and 100 bp larger than the expected band. Libraries were sequenced at the UCI Genomics High-throughput Facility on an Illumina MiSeq using the 500-cycle v2 reagent kit (Illumina Cat # MS-102-2003).

Amplification bias assay of varying polymerases. Since the bias for TdT-mediated insertions is for the nucleotides G or C, it was important to test which polymerase would be optimal for amplifying GC stretches inserted on the hgRNA. Three gBlock Gene Fragments (IDT) were ordered with an insertion of 40 nucleotides at the -3 position of the hgRNA. The insertion was either 50%, 65%, or 80% GC rich. The amplification test was either performed with Q5 Hot Start High-Fidelity DNA Polymerase or Phusion Hot Start Flex DNA Polymerase (NEB). Reactions were performed with 10 ng of the individual fragments as well as with a 1:1:1 mix of all three fragments. Both forward and reverse primers contained a 5 – 7 nt sample-specific barcode and the appropriate Illumina adapters. The eight PCR's were performed with the following protocol: 98 °C, 3 min; (98 °C, 1 min; 55 °C, 30 s; 72 °C, 30 s) × 35; 72 °C, 1 min with a 2 °C/s ramp rate. Each PCR was normalized using agarose gel electrophoresis and equal amounts of each sample were pooled based on signal intensity. Final pools were cleaned on a NucleoSpin Gel and PCR Clean-up Kit column (Macherey-Nagel) and sent to Quintara Biosciences or FornaxBio and sequenced as above.

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) and mapped to the reference sequence by the alignment algorithm implementation used in Perli et al, 2016 which provides a sequence of M(Match), X(Mismatch), I(Insertion) and D(Deletion) as the mapping result.

If UMI barcodes are present, before mapping, sequences with the same barcodes are combined to one. To combine, we started with a multiple-alignment of the sequences (done by Motility library in Python) with the same UMI barcode (One nucleotide difference was allowed in UMI barcodes) and to avoid any random mismatches produced in the sequencing process, for each position in this alignment only nucleotides present in more than 50% of the sequences in this group, were used to generate the consensus sequence.

After the alignment, first bad alignments (>20 mismatches or >50% deletions) were removed. Then, mismatches and inserted or deleted sequences, their positions on the reference sequence and their frequencies were extracted. Only insertions or deletions occurring around (-10nt to +10nt) the cut side were kept. (For the data in Figures 2 and S2, we used the region -7 bp to +7 bp from the cut site to avoid a genomic SNP.) 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 the plasmids that were transfected, they were filtered out. In addition, insertions longer than 15 nt (or 20 nt for Figure 2C and 40 nt for S4) 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 on Gitlab at https://gitlab.com/__mason__/CHRYON.

For the experiment shown in Figures 5, S5, and S6, due to the shorter read length, forward and reverse reads were not paired, and forward reads only were used for the analysis.

Western blot. For determining protein expression of our Cas9 constructs, western blots were performed by first lysing cell pellets with 1X RIPA buffer and a protease inhibitor cocktail (Roche #4693159001). After 30 minutes on ice, the lysis reaction was spun down and the supernatant was used in a BCA Reagent Assay (ThermoFisher Scientific #PI23225) to normalize for protein concentration. Upon normalization, the necessary volume of supernatant was added to LDS Sample Buffer (ThermoFisher Scientific #NP0007) with 0.2% βME (Fisher Scientific #BP176100).

All protein gels were 4-12% Bis-Tris Gels, 1.0 mm, 15-well (Invitrogen, Thermo Scientific #NP0323) and electrophoresis was performed in an XCell SureLock Mini-Cell Electrophoresis System (ThermoFisher Scientific #EI0001) with 1X MOPS Running Buffer according to the NuPAGE MOPS SDS Running Buffer recipe. Protein transfers were performed in a Mini Trans-Blot Cell (Bio-Rad #1703930) with a transfer buffer made according to the Bjerrum Schafer-Nielsen Buffer with SDS (Bio-Rad). Protein membranes and blotting paper were components of the EMD Millipore Blotting Sandwich Immobilin-P (Millipore Sigma #IPSN07852).

After transfer, the protein membrane was cut to create separate sections for Cas9 and actin blotting. An additional protein gel and transfer were performed for experiments that blotted for TdT. The respective membrane was incubated in either Guide-it Cas9 Polyclonal Antibody (Clontech #632607), α-TdT Antibody (Abcam #ab14772), or α-Actin Antibody (Abcam #ab14128) for 3 hours at room temperature. Western blots were then incubated with horseradish peroxidase-fused secondary antibodies. α-Rabbit IgG (Sigma-Aldrich #A0545) was used to bind the primary antibody of Cas9 and TdT, while α-Mouse IgG (R&D Systems #HAF007) was used to bind the primary antibody of actin. The western blot membrane was treated with Clarity ECL Western Blotting Substrate (Bio-Rad #1705061). Blots were scanned on a ChemiDoc Touch Imaging System (Bio-Rad #1708370).

References

Bhan, N., Strutz, J., Glaser, J. I., Kalhor, R., Boyden, E., Church, G. M., et al. (2019). Recording temporal data onto DNA with minutes resolution, 1–13. http://doi.org/10.1101/634790

Campeau, E., Ruhl, V. E., Rodier, F., Smith, C. L., Rahmberg, B. L., Fuss, J. O., et al. (2009). A versatile viral system for expression and depletion of proteins in mammalian cells. *PLoS ONE*, *4*(8), e6529. http://doi.org/10.1371/journal.pone.0006529

Chan, M. M., Smith, Z. D., Grosswendt, S., Kretzmer, H., Norman, T. M., Adamson, B., et al. (2019). Molecular recording of mammalian embryogenesis. *Nature*, 1–23. http://doi.org/10.1038/s41586-019-1184-5

Ede, C., Chen, X., Lin, M.-Y., & Chen, Y. Y. (2016). Quantitative Analyses of Core Promoters Enable Precise Engineering of Regulated Gene Expression in Mammalian Cells. *ACS Synthetic Biology*, *5*(5), 395–404. http://doi.org/10.1021/acssynbio.5b00266

Frieda, K. L., Linton, J. M., Hormoz, S., Choi, J., Chow, K.-H. K., Singer, Z. S., et al. (2017). Synthetic recording and in situ readout of lineage information in single cells. *Nature*, *541*(7635), 107–111. http://doi.org/10.1038/nature20777

Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M., & Joung, J. K. (2014). Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nature Biotechnology*, *32*(3), 279–284. http://doi.org/10.1038/nbt.2808

Gisler, S., Gonçalves, J. P., Akhtar, W., de Jong, J., Pindyurin, A. V., Wessels, L. F. A., & van Lohuizen, M. (2019). Multiplexed Cas9 targeting reveals genomic location effects and gRNA-based staggered breaks influencing mutation efficiency. *Nature Communications*, *10*(1), 1598. http://doi.org/10.1038/s41467-019-09551-w

Glaser, J. I., Zamft, B. M., Marblestone, A. H., Moffitt, J. R., Tyo, K., Boyden, E. S., et al. (2013). Statistical analysis of molecular signal recording. *PLoS Computational Biology*, *9*(7), e1003145. http://doi.org/10.1371/journal.pcbi.1003145

Hwang, B., Lee, W., Yum, S.-Y., Jeon, Y., Cho, N., Jang, G., & Bang, D. (2019). Lineage tracing using a Cas9-deaminase barcoding system targeting endogenous L1 elements. *Nature Communications*, *10*(1), 1234. http://doi.org/10.1038/s41467-019-09203-z

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*(6096), 816–821. http://doi.org/10.1126/science.1225829

Junker, J. P., Spanjaard, B., Peterson-Maduro, J., Alemany, A., Hu, B., Florescu, M., & van Oudenaarden, A. (2016). Massively parallel whole-organism lineage tracing using CRISPR/Cas9 induced genetic scars. *bioRxiv*, 056499. http://doi.org/10.1101/056499

Kalhor, R., Kalhor, K., Mejia, L., Leeper, K., Graveline, A., Mali, P., & Church, G. M. (2018). Developmental barcoding of whole mouse via homing CRISPR. *Science*, *361*(6405), eaat9804. http://doi.org/10.1126/science.aat9804

Kalhor, R., Mali, P., & Church, G. M. (2017). Rapidly evolving homing CRISPR barcodes. *Nature Methods*, *14*(2), 200. http://doi.org/10.1038/nmeth.4108

Kleinstiver, B. P., Prew, M. S., Tsai, S. Q., Topkar, V. V., Nguyen, N. T., Zheng, Z., et al. (2015). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*, *523*(7561), 481–485. http://doi.org/10.1038/nature14592

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., & Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, *533*(7603), 420–424. http://doi.org/10.1038/nature17946

Landau, N. R., Schatz, D. G., Rosa, M., & Baltimore, D. (1987). Increased frequency of N-region insertion in a murine pre-B-cell line infected with a terminal deoxynucleotidyl transferase retroviral expression vector. *Molecular and Cellular Biology*, *7*(9), 3237–3243. http://doi.org/10.1128/mcb.7.9.3237

Lee, H. H., Kalhor, R., Goela, N., Bolot, J., & Church, G. M. (2018). Enzymatic DNA synthesis for digital information storage, 1–31. http://doi.org/10.1101/348987

Liu, M., Maurano, M. T., Wang, H., Qi, H., Song, C.-Z., Navas, P. A., et al. (2015). Genomic discovery of potent chromatin insulators for human gene therapy. *Nature Biotechnology*, *33*(2), 198–203. http://doi.org/10.1038/nbt.3062

Mahajan, K. N., Gangi-Peterson, L., Sorscher, D. H., Wang, J., Gathy, K. N., Mahajan, N. P., et al. (1999). Association of terminal deoxynucleotidyl transferase with Ku. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(24), 13926–13931. http://doi.org/10.1073/pnas.96.24.13926

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., et al. (2013). RNA-guided human genome engineering via Cas9. *Science*, *339*(6121), 823–826. http://doi.org/10.1126/science.1232033

McDole, K., Guignard, L., Amat, F., Berger, A., Malandain, G., Royer, L. A., et al. (2018). In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level. *Cell*, *175*(3), 859–876.e33. http://doi.org/10.1016/j.cell.2018.09.031

McKenna, A., Findlay, G. M., Gagnon, J. A., Horwitz, M. S., Schier, A. F., & Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, *353*(6298), aaf7907. http://doi.org/10.1126/science.aaf7907

Motea, E. A., & Berdis, A. J. (2010). Terminal deoxynucleotidyl transferase: The story of a misguided DNA polymerase. *Biochimica Et Biophysica Acta (BBA) - Proteins and Proteomics*, *1804*(5), 1151. http://doi.org/10.1016/j.bbapap.2009.06.030

Palluk, S., Arlow, D. H., de Rond, T., Barthel, S., Kang, J. S., Bector, R., et al. (2018). De novo DNA synthesis using polymerase-nucleotide conjugates. *Nature Biotechnology*, *36*(7), 645–650. http://doi.org/10.1038/nbt.4173

Perli, S. D., Cui, C. H., & Lu, T. K. (2016). Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science*, *353*(6304), aag0511–aag0511. http://doi.org/10.1126/science.aag0511

Pryor, J. M., Conlin, M. P., Carvajal-Garcia, J., Luedeman, M. E., Luthman, A. J., Small, G. W., & Ramsden, D. A. (2018). Ribonucleotide incorporation enables repair of chromosome breaks by nonhomologous end joining. *Science*, *361*(6407), 1126–1129. http://doi.org/10.1126/science.aat2477

Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., et al. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology*, *36*(5), 442–450. http://doi.org/10.1038/nbt.4103

Rankin, E. B., & Giaccia, A. J. (2016). Hypoxic control of metastasis. *Science*, *352*(6282), 175–180. http://doi.org/10.1126/science.aaf4405

Repasky, J. A. E., Corbett, E., Boboila, C., & Schatz, D. G. (2004). Mutational analysis of terminal deoxynucleotidyltransferase-mediated N-nucleotide addition in V(D)J recombination. *Journal of Immunology (Baltimore, Md. : 1950)*, *172*(9), 5478–5488.

Schmidt, S. T., Zimmerman, S. M., Wang, J., Kim, S. K., & Quake, S. R. (2017). Quantitative Analysis of Synthetic Cell Lineage Tracing Using Nuclease Barcoding. *ACS Synthetic Biology*, *6*(6), 936–942. http://doi.org/10.1021/acssynbio.6b00309

Semenza, G. L. (2012). Hypoxia-Inducible Factors in Physiology and Medicine. *Cell*, *148*(3), 399–408. http://doi.org/10.1016/j.cell.2012.01.021

Sheth, R. U., & Wang, H. H. (2018). DNA-based memory devices for recording cellular events. *Nature Reviews. Genetics*, *19*(11), 718–732. http://doi.org/10.1038/s41576-018-0052-8

Sheth, R. U., Yim, S. S., Wu, F. L., & Wang, H. H. (2017). Multiplex recording of cellular events over time on CRISPR biological tape. *Science*, *358*(6369), 1457–1461. http://doi.org/10.1126/science.aao0958

Shipman, S. L., Nivala, J., Macklis, J. D., & Church, G. M. (2016). Molecular recordings by directed CRISPR spacer acquisition. *Science*, *353*(6298), aaf1175. http://doi.org/10.1126/science.aaf1175

Shipman, S. L., Nivala, J., Macklis, J. D., & Church, G. M. (2017). CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature*, *547*(7663), 345–349. http://doi.org/10.1038/nature23017

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., & Junker, J. P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nature Biotechnology*, *36*(5), 469–473. http://doi.org/10.1038/nbt.4124

Tang, W., & Liu, D. R. (2018). Rewritable multi-event analog recording in bacterial and mammalian cells. *Science*, *360*(6385), eaap8992. http://doi.org/10.1126/science.aap8992

Tsai, S. Q., Wyvekens, N., Khayter, C., Foden, J. A., Thapar, V., Reyon, D., et al. (2014). Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nature Biotechnology*, *32*(6), 569–576. http://doi.org/10.1038/nbt.2908

Waldo, G. S., Standish, B. M., Berendzen, J., & Terwilliger, T. C. (1999). Rapid protein-folding assay using green fluorescent protein. *Nature Biotechnology*, *17*(7), 691–695. http://doi.org/10.1038/10904

Wang, T., Wei, J. J., Sabatini, D. M., & Lander, E. S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, *343*(6166), 80–84. http://doi.org/10.1126/science.1246981

Yan, Q., Bartz, S., Mao, M., Li, L., & Kaelin, W. G. (2007). The hypoxia-inducible factor 2alpha N-terminal and C-terminal transactivation domains cooperate to promote renal tumorigenesis in vivo. *Molecular and Cellular Biology*, *27*(6), 2092–2102. http://doi.org/10.1128/MCB.01514-06

Yang, B., Gathy, K. N., & Coleman, M. S. (1994). Mutational analysis of residues in the nucleotide binding domain of human terminal deoxynucleotidyl transferase. *Journal of Biological Chemistry*, *269*(16), 11859–11868.

Zamft, B. M., Marblestone, A. H., Kording, K., Schmidt, D., Martin-Alarcon, D., Tyo, K., et al. (2012). Measuring cation dependent DNA polymerase fidelity landscapes by deep sequencing. *PLoS ONE*, *7*(8), e43876. http://doi.org/10.1371/journal.pone.0043876

Zuo, Z., & Liu, J. (2016). Cas9-catalyzed DNA Cleavage Generates Staggered Ends: Evidence from Molecular Dynamics Simulations. *Scientific Reports*, *5*(1), 37584. http://doi.org/10.1038/srep37584

Acknowledgements: We thank Seanjeet K. Paul and Tate C. Lone for technical assistance and all members of the Liu Laboratory and Chengjian Li for helpful discussions. We thank the following people for plasmids: Yvonne Chen, Keith Joung, William Kaelin, George Church, David Liu, Eric Campeau, Paul Kaufman, and Timothy Lu. This work was made possible, in part, through access to the Genomics High Throughput Facility Shared Resource of the Cancer Center Support Grant (P30CA-062203) at the University of California, Irvine and NIH shared instrumentation grants 1S10RR025496-01, 1S10OD010794-01, and 1S10OD021718-01. This work was funded by NIH grants 1DP2GM119163-01 and 1R21GM126287-01 to CCL, an AHA Predoctoral Fellowship to CKC, and a fellowship from the NSF-Simons Center for Multiscale Cell Fate Research (NSF Award #1763272) to TBL.

Author contributions: TBL and CCL designed experiments. TBL, JHG, MWS, and CKC performed experiments. TBL, JHG, GL, and CCL developed protocols. TBL, JHG, and CKC made cell lines. TBL, JHG, CKC, and GL made and validated reagents. TBL, MWS, EF, BSA, XX, and CCL designed and performed analyses. MWS, EF, and BSA wrote code. TBL and CCL wrote the paper, with input from all authors. CCL procured funding and oversaw the project.

Availability of data and reagents: All NGS data sets will be deposited at the NCBI's Sequence Read Archive. All plasmids will be available soon at Addgene. Please contact CCL if you would like a plasmid before it is available at Addgene. See **Supplemental Table S6** for a guide to these reagents. Please contact CCL for cell lines.

Supplemental Material Contents:

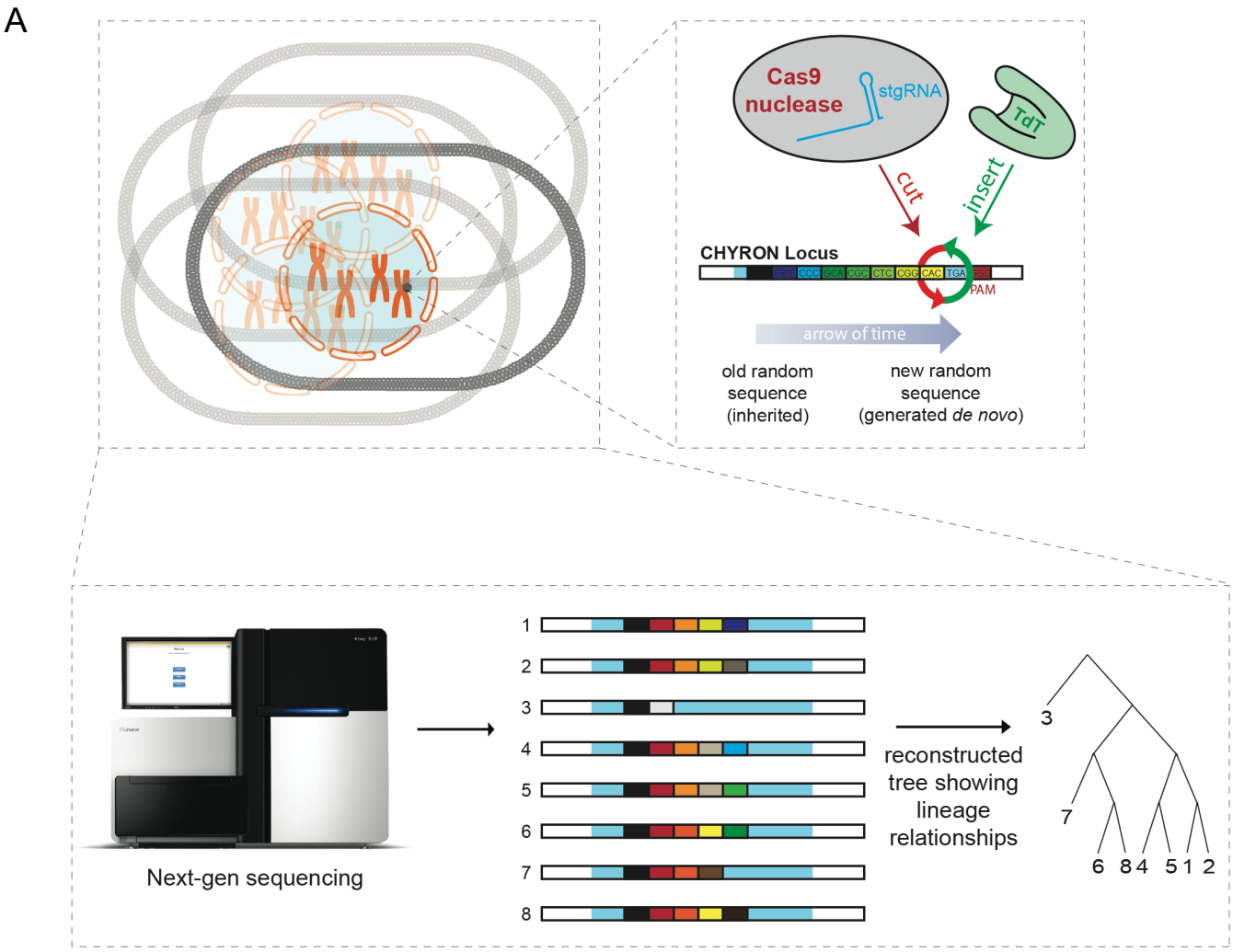
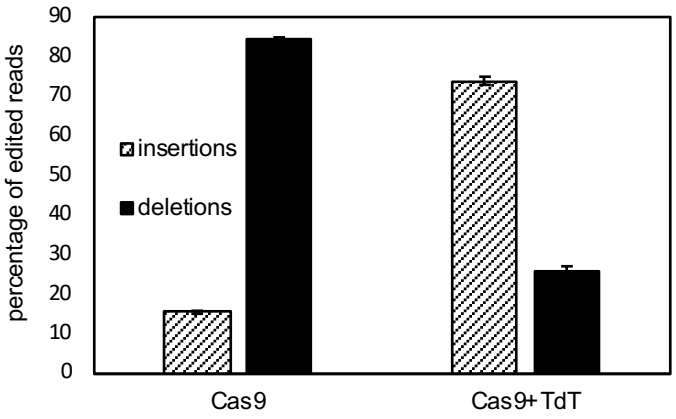
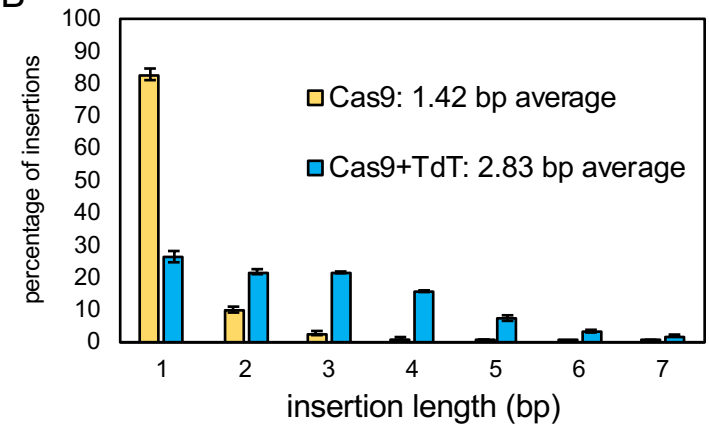


Figure 1. Design of the CHYRON locus. (A) The constitutive expression of Cas9 and TdT mediates the ordered acquisition of random insertion mutations (represented as differently-colored boxes). Each cell is represented by one unique sequence, which can be compared to other sequences to reconstruct the lineage of the cells. **(B)** If Cas9 and TdT expression is tied to a stimulus, the lengths of the acquired insertions can be interpreted to determine each cell's relative dose of the stimulus. **(C)** A homing or self-targeting guide RNA, which directs Cas9 to the DNA that encodes it.

A



B



C

protospacer -4/-3 nts	A	C	G	T
AA*	15.9	25.7	47.0	11.3
AC**	16.6	26.0	46.3	11.0
AG	23.8	42.7	22.7	10.8
AT	12.6	39.7	34.7	13.0
CA	16.1	31.4	37.9	14.6
CC	13.1	41.4	25.9	19.7
CG	13.9	41.0	30.1	15.0
CT*	7.9	49.2	16.8	26.2
GA	9.3	36.8	34.7	19.2
GC	16.7	40.2	27.6	15.5
GG	24.0	18.0	47.0	11.0
GT	6.6	53.1	18.5	21.8
TA	12.5	35.6	38.4	13.5
TC	17.3	30.6	36.5	15.6
TG*	13.9	38.7	28.5	18.9
TT*	17.6	35.3	34.0	13.2

D

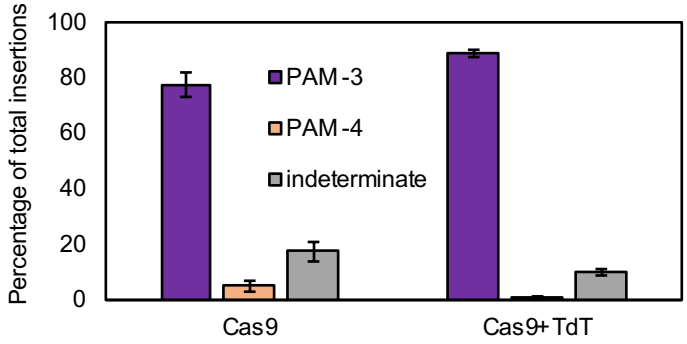


Figure 2.

Figure 2. Expression of TdT promotes random insertions at a Cas9-induced DSB. (A) Expression of TdT promotes insertion mutations. 293T cells were transfected with plasmids expressing Cas9 and TdT, or Cas9 alone, and an sgRNA against a genomic site (HEK293site3). Three days later, cells were collected, DNA was extracted, and the targeted genomic site was amplified by PCR and sequenced by NGS. Bars represent the mean of six replicates (two technical replicates each of three biological replicates). Error bars = ± stdev. Sequences were annotated as unchanged, pure insertions (insertions), or any sequence that leads to a loss of information (deletions). "Deletions" included pure deletions, mixtures of insertion, deletion, and substitution mutations, and a small proportion (2% of all edited sequences) of pure substitutions. **(B)** Expression of TdT creates longer insertion mutations than those created in the presence of Cas9 alone. Of the pool of pure insertions, the percentage of each length were calculated and plotted. Bars represent the mean of the six replicates. Error bars = ± stdev. **(C)** Insertion sequences are random, but have a bias toward G and C nucleotides. 293T cells were transfected with plasmids expressing Cas9 and TdT, and one of 16 sgRNAs against different genomic sites. The target protospacers were chosen to have all possible combinations of nucleotides at the sites 4 and 3 nt upstream of the PAM sequence on the top (non-target) strand. The proportions of each nucleotide (on the top strand) found in all pure insertion sequences 4 bp in length were calculated for each protospacer. Data shown are the average of four replicates (two technical replicates each of two biological replicates), except those marked with *, which are the average of two technical replicates of a single biological replicate, and the row marked with **, which are the average of two biological replicates. **(D)** TdT-mediated insertions are added 3 bp upstream of the PAM. For the pure insertions show in (A), the position of the insertion was calculated, if the insertion sequence made this calculation possible. An Insertion would be annotated as having an "indeterminate" position, for example, if the 3' nt of the insertion were identical to the protospacer nt 5' of where the insertion was placed.

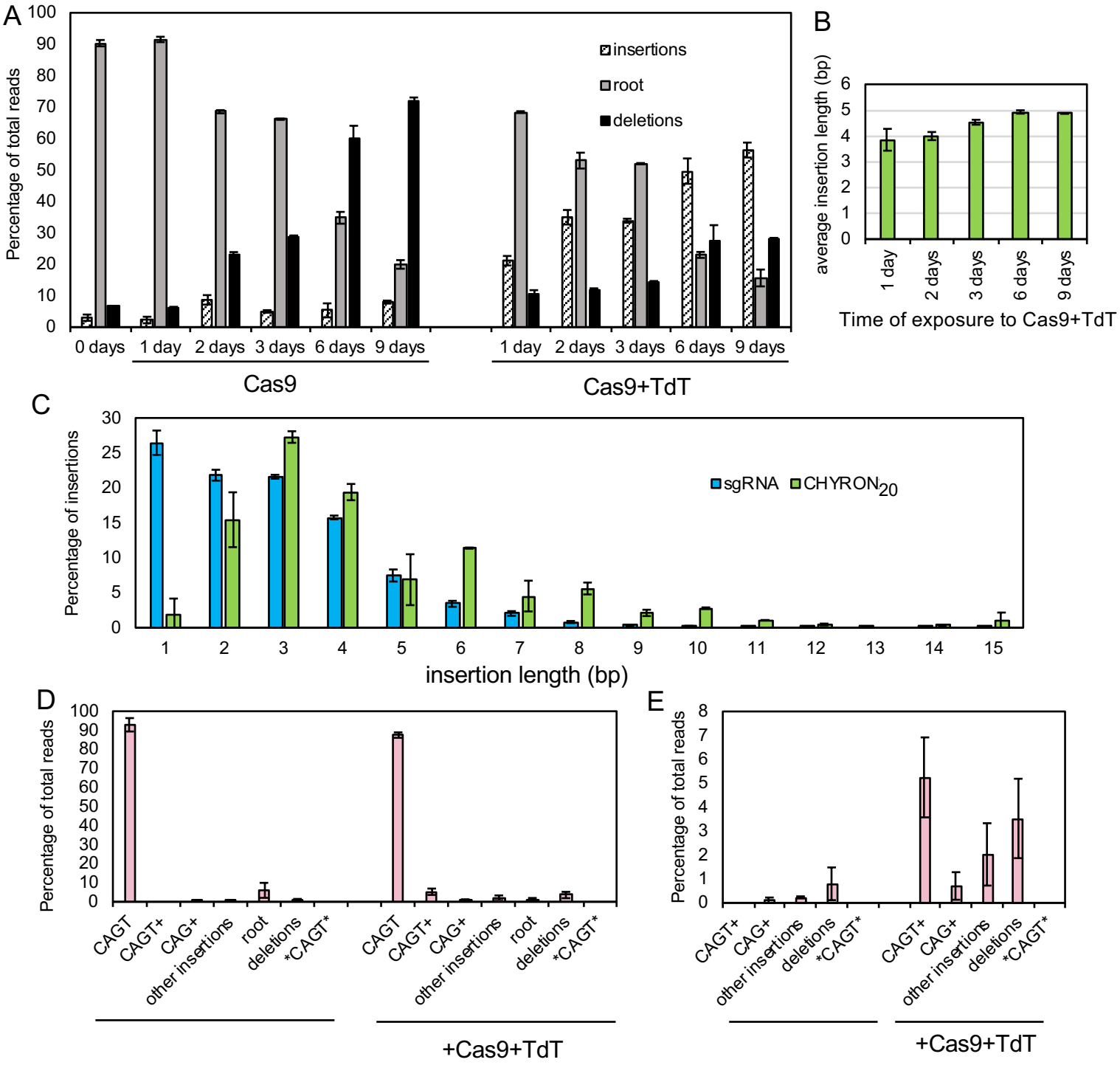


Figure 3.

Figure 3. An integrated hgRNA accumulates insertions in multiple rounds. A clonal 293T cell line bearing an integrated hgRNA (hereafter 293T-CHYRON$_{20}$) was created so that expression of Cas9 and TdT leads to multiple rounds of insertion of random nucleotides in an ordered fashion. **(A)** 293T-CHYRON$_{20}$ cells accumulate insertions and deletions over time when exposed to Cas9 and TdT. 293T-CHYRON$_{20}$ cells were transfected with a plasmid expressing Cas9 and, optionally, TdT for the indicated time before collection. Cells were re-transfected every three days. The hgRNA locus was sequenced and each sequence was annotated as unchanged (root), pure insertion (insertions), or any sequence that involves a loss of information (deletions). Bars represent the average of two technical replicates. Error bars=$\pm$stdev. **(B)** Insertions grow longer over time, until the six day timepoint, then stop growing. The average length of pure insertions at each timepoint was calculated. **(C)** The distribution of insertion lengths is longer for an hgRNA than for a protospacer targeted in a single round. Insertion lengths after 3 days were compared to insertions at a genomic site with the same spacer sequence targeted with an sgRNA (data from Figure 2B). **(D)** Cas9 and TdT mediate multiple rounds of editing on an integrated hgRNA. The 293T-CHYRON$_{20}$ cell line was transfected with Cas9 and TdT to induce insertions, then purified to near-clonality. This near-clonal cell line bearing an insertion with the sequence CAGT was then transfected again with a plasmid expressing Cas9 and TdT. These cells, and an untransfected control, were grown for 6-15 days, then collected. The CHYRON locus was sequenced and editing outcomes were determined to be the root CHYRON$_{20}$ sequence (root), deletions, a CAGT insertion (CAGT), an insertion containing the prefix CAGT or CAG (CAGT+ or CAG+, respectively), an insertion containing the sequence CAGT other than as a prefix (*CAGT*), or other insertions. Error bars=$\pm$stdev of two technical replicates each of the 6 and 15 day timepoints. **(E)** Replotting of some data from (D) for easier comparison.

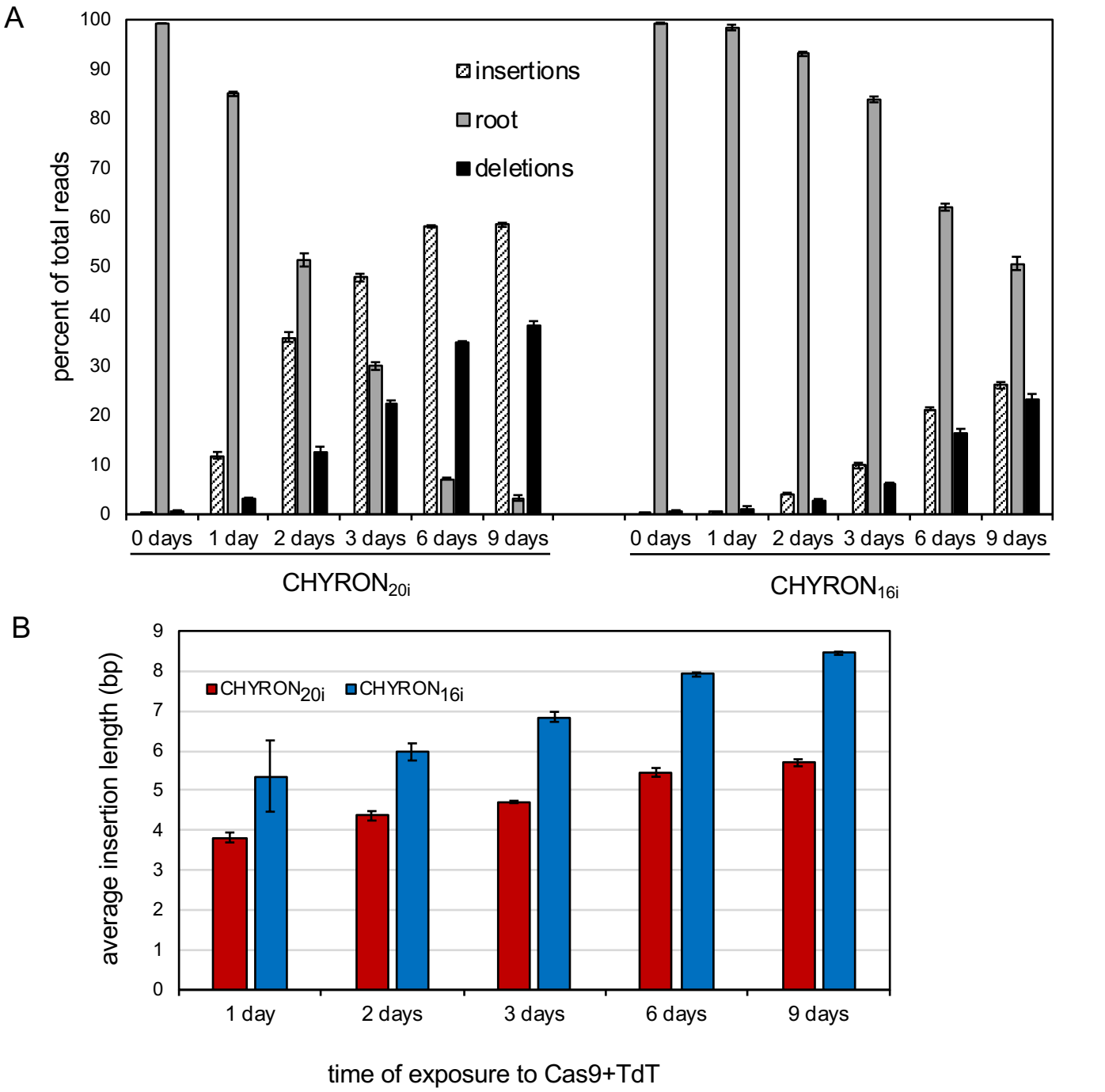
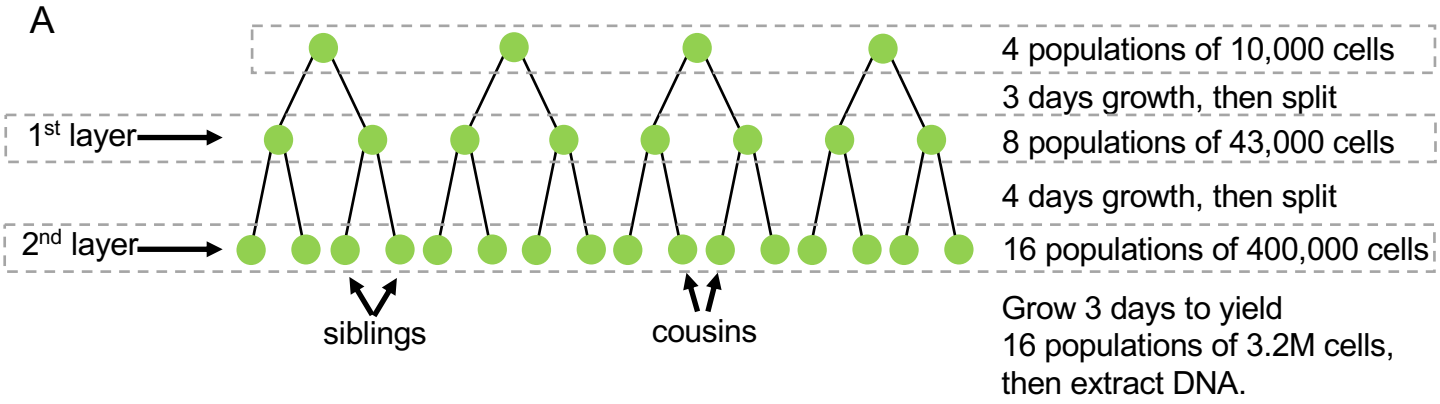
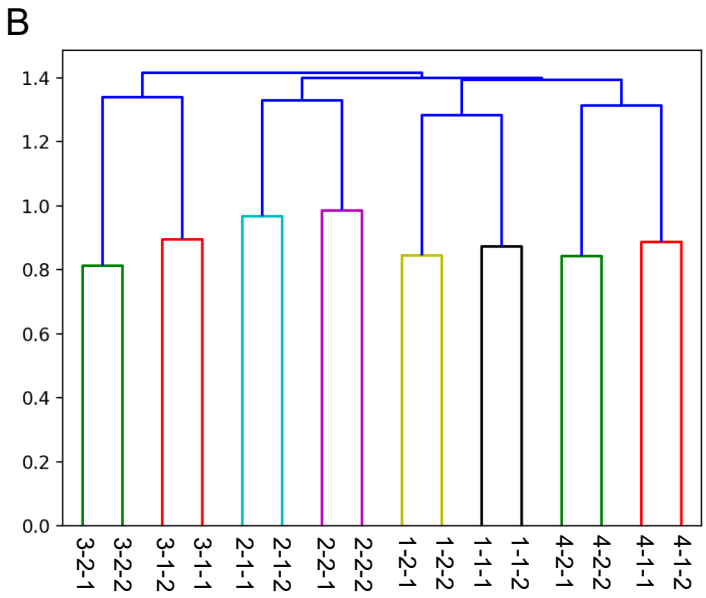


Figure 4. An improved version of CHYRON accumulates insertions of average length 8.4 in an average of three rounds. (A) CHYRON loci with an initial hgRNA length of 20 nt accumulate insertions and deletions over six days, whereas those with an initial hgRNA length of 16 nt accumulate insertions and deletions more slowly, continuing to do so through a 9-day timecourse. Clonal 293T cell lines (hereafter 293T-CHYRON$_{20i}$ and 293T-CHYRON$_{16i}$), were created by integrating cassettes at the *AAVS1* safe harbor locus. The cassettes contain hgRNAs with initial lengths of 20 or 16 nt, flanked by insulator sequences to prevent silencing. 293T-CHYRON$_{20i}$ and 293T-CHYRON$_{16i}$ cells were transfected with a plasmid expressing Cas9 and TdT for the indicated time before collection. Cells were re-transfected every three days. The CHYRON locus was analyzed by NGS and each sequence was annotated as root, pure insertion (insertions) or any sequence that involves a loss of information (deletions). Bars represent the average of three technical replicates. Error bars=±stdev. **(B)** Insertions continue to grow throughout the 9-day timecourse. The average lengths of pure insertions were calculated from the experiment in A. Bars represent the average of three technical replicates. Error bars=±stdev.


A
4 populations of 10,000 cells
3 days growth, then split
1st layer
8 populations of 43,000 cells
4 days growth, then split
2nd layer
16 populations of 400,000 cells
siblings
cousins
Grow 3 days to yield
16 populations of 3.2M cells,
then extract DNA.


B
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
3-2-1
3-2-2
3-1-2
3-1-1
2-1-1
2-1-2
2-2-1
2-2-2
1-2-1
1-2-2
1-1-1
1-1-2
4-2-1
4-2-2
4-1-1
4-1-2

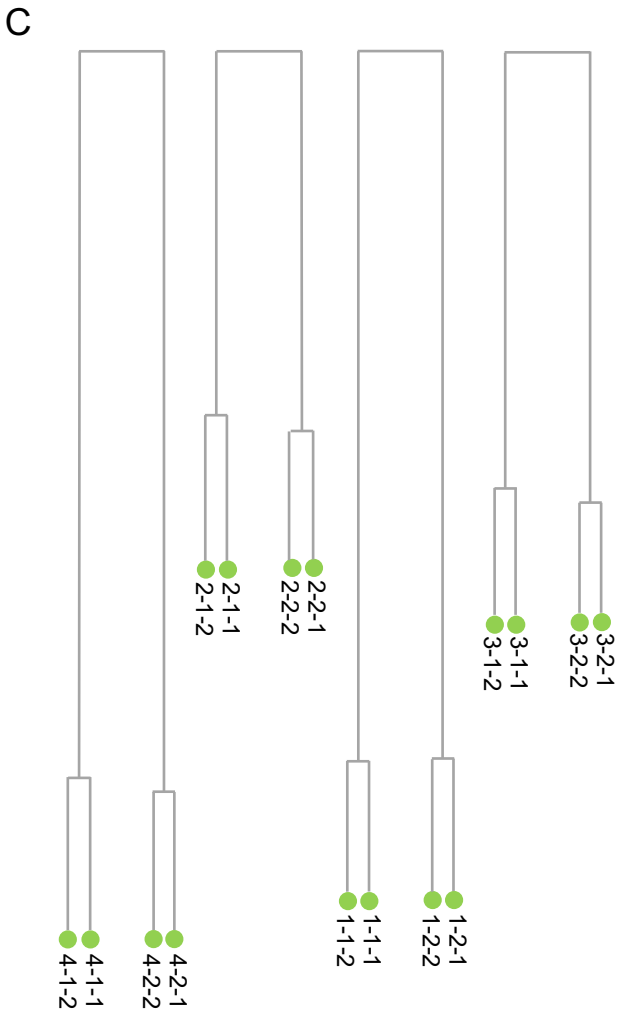

C
4-1-2
4-1-1
4-2-2
4-2-1
2-1-2
2-1-1
2-2-2
2-2-1
1-1-2
1-1-1
1-2-2
1-2-1
3-1-2
3-1-1
3-2-2
3-2-1

Figure 5.

Figure 5. Reconstruction of cell relatedness by sequencing of the CHYRON locus.
(A) Plan of the experiment. 5,000 293T-CHYRON$_{16i}$ cells were plated in each of 4 wells of a 384-well plate, then transfected the next day with a plasmid expressing Cas9 and TdT, then each well was moved to one well of a 96-well plate the next day. Three days after transfection, when they had expanded to approximately 86,000 cells per well, they were each split into two wells of a 24-well plate, allowed to attach for one day, then transfected again. Three days later, when each well had expanded to approximately 800,000 cells, each well was split into two wells of a 6-well dish. Three days later, all wells were collected and sequenced. **(B)** A simple method leads to perfect reconstruction of the relatedness of all wells. For each well, a list was created of all unique insertions with an abundance of at least 0.0139% of the non-deletion reads and a length of 8-15 bp. This list was used to create a binary vector for each well. The length of the vector was equivalent to the number of insertions in all wells that met the abundance and length criteria. In each well's vector, insertions that are present in that well are indicated with a 1 and absent insertions are indicated with a 0. These vectors were used to compute the Jaccard distances between all pairs of wells. Then, hierarchical reconstruction was performed using the UPGMA algorithm. **(C)** The relatedness of all wells was perfectly reconstructed using a new method that makes use of the ordered nature of insertions at the CHYRON locus. An "average prefix" was calculated for all possible pairs of wells. To calculate the average prefix between wells A and B, for example, for each insertion in well A, the insertion in well B with the most initial nucleotides in common (the "longest prefix") is found. For that pairing, the number of initial nucleotides shared between the two sequences (the "prefix length") is recorded. This process is repeated for every other insertion in well A, then for every insertion in well B. The prefix lengths of all of the paired insertions are then averaged to calculate the "average prefix" between the two wells. Once the average prefix for each pair of wells has been calculated, the two wells with the longest average prefix are paired, then the remaining wells with the longest average prefix are paired, and so on until all wells are paired. Thus, all pairs of siblings are identified. Then, for each pair of siblings, a "pooled well" was created, consisting of insertions common to each well. The average prefix process was then repeated on these pooled wells. The distance between each pair of wells was visualized by determining the average prefix of insertions in the two wells, subtracting the average prefix of insertions in the two most-distantly related wells, then inverting the result and setting all the distances proportionally.

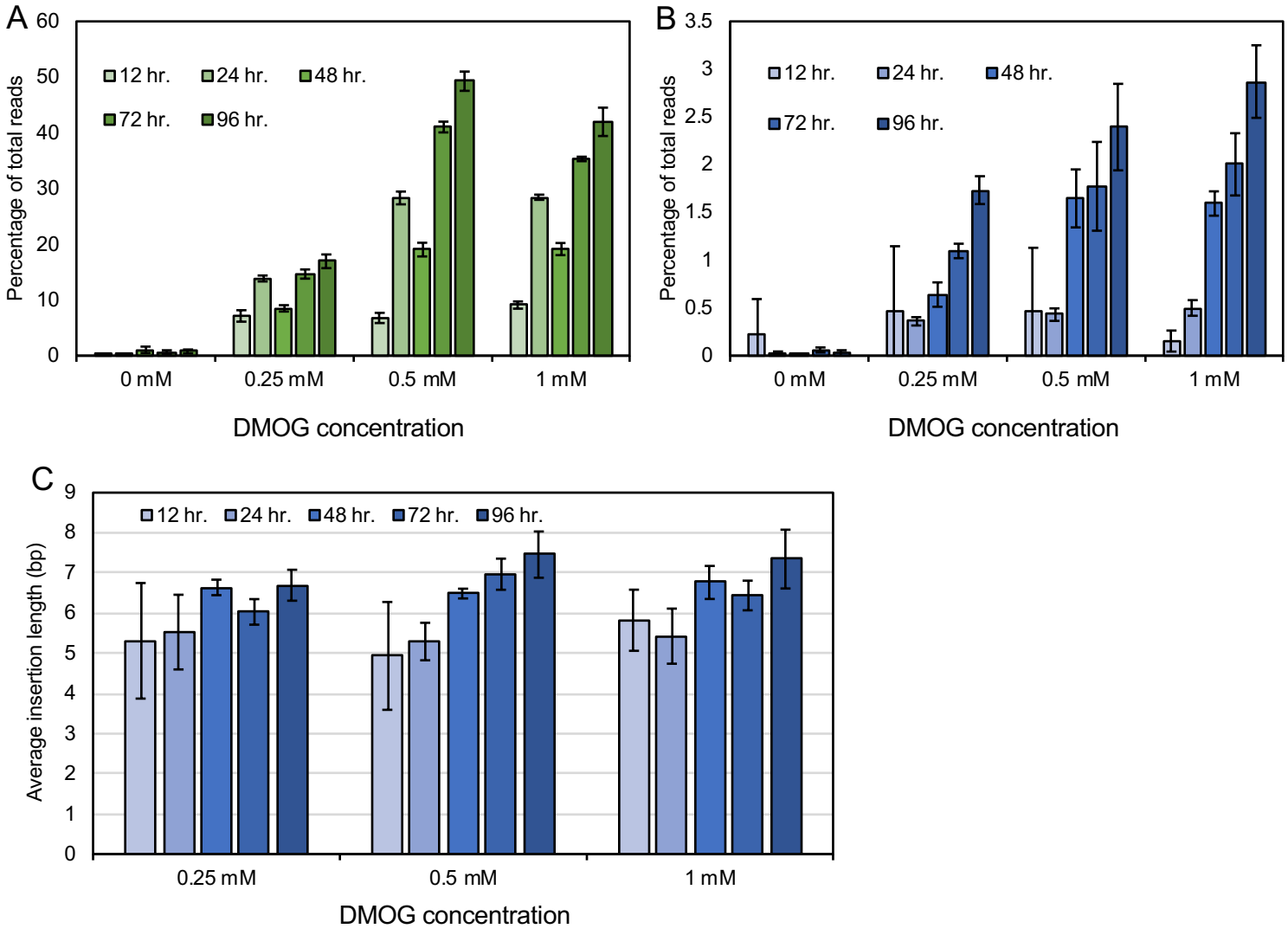


Figure 6. Expression of hypoxia-inducible Cas9 and TdT promotes insertion abundance and length in proportion to duration of treatment with the hypoxia mimic DMOG. **(A)** When transfected with hypoxia-inducible Cas9 and TdT, an ever-higher proportion of $CHYRON_{20}$ loci accumulate insertions upon longer treatment and higher doses of DMOG, a hypoxia mimic. 293T-$CHYRON_{20}$ cells were transfected with a plasmid encoding Cas9 and TdT under the control of a promoter containing four copies of the hypoxia response element; Cas9 is additionally fused to a degron that destabilizes proteins in the presence of normal levels of oxygen. After transfection, cells were treated with the DMOG or a vehicle control and then collected at the indicated time and analyzed as in Figure 3A. Bars represent the average of three technical replicates. Error bars=±stdev. **(B)** CHYRON16 loci accumulate insertions at a lower, but still dose-dependent rate when transfected with hypoxia-inducible Cas9 and TdT and exposed to DMOG. 293T-$CHYRON_{16}$ cells were transfected and insertions analyzed as in (A). **(C)** In 293T-CHYRON16 cells transfected with hypoxia-inducible Cas9 and TdT, insertions grow longer with increasing duration of exposure to DMOG. Lengths of pure insertions in the $CHYRON_{16}$ experiment were calculated. Bars represent the average of three technical replicates. Error bars=±stdev.