nature chemical biology

ARTICLES

https://doi.org/10.1038/s41589-021-00769-8

Check for updates

Lineage tracing and analog recording in mammalian cells by single-site DNA writing

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

Studying cellular and developmental processes in complex multicellular organisms can require the non-destructive observation of thousands to billions of cells deep within an animal. DNA recorders address the staggering difficulty of this task by converting transient cellular experiences into mutations at defined genomic sites that can be sequenced later in high throughput. However, existing recorders act primarily by erasing DNA. This is problematic because, in the limit of progressive erasure, no record remains. We present a DNA recorder called CHYRON (Cell History Recording by Ordered Insertion) that acts primarily by writing new DNA through the repeated insertion of random nucleotides at a single locus in temporal order. To achieve in vivo DNA writing, CHYRON combines Cas9, a homing guide RNA and the template-independent DNA polymerase terminal deoxynucleotidyl transferase. We successfully applied CHYRON as an evolving lineage tracer and as a recorder of user-selected cellular stimuli.

Observation of living organisms as they develop 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 with advances in microscopy and the continuing development of genetically encoded labels that can be imaged non-destructively (for example, GFP). However, live imaging of single cells in intact organisms is still constrained by context and scale. Animals, for example, tend to be opaque, and, even when developmental processes are accessible to microscopy[1], cell tracking poses substantial computational and data management challenges when the number of cells under observation exceeds tens of thousands. An alternative paradigm to real-time observation is DNA recording. In DNA recording, transient cellular events are engineered to trigger permanent mutations in a cell's own genome (Fig. 1a,b). As DNA is both durable and propagating and 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 and read out later at unprecedented depth. Although the reading step is destructive, recording is not, creating an effective alternative to real-time observation that can scale to millions of cells in opaque model animals such as mice.

DNA recording has recently been transformed by the development of genetically encoded CRISPR-based systems that drive the rapid accumulation of mutations at neutral loci in cellular genomes[2–17]. 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[3,5,7,8]; when activity is constitutive, accumulated mutations capture lineage relationships between individual cells[2,4,6,11–13,16–18]. Two recent architectures for CRISPR-based recording systems are particularly amenable to recording at a single genomic locus in large numbers of mammalian cells. The first architecture relies on arrays of *Streptococcus pyogenes* Cas9 target sites[2,12,13,17,18]. Here, Cas9 targets random elements of the array to generate insertions and deletions (indels) at array elements, such that the progressive accumulation of indels across the array marks cells with their lineage relationships or their history of exposure to a stimulus. The second architecture relies on a self-targeting[3] or homing[4,11] guide RNA (hgRNA) that directs Cas9 to the very locus from which the hgRNA is expressed (Fig. 1c). The locus changes over time through a series of indels for which the pattern reflects lineage information[4,11] or exposure to stimuli[3]. These two types of DNA recording systems were used to identify the early and late embryonic origin of thousands of cell lineages in adult zebrafish[2,12,13,18], trace hematopoiesis in mice[17] and record inflammation exposure in HEK293T cells implanted into mice treated with lipopolysaccharide[3]. However, as the progressive accumulation of deletions at a single site quickly corrupts or removes previous deletion patterns[3,4], and continuous editing of an array of sites at a single locus can lead to multiple simultaneous cuts with loss of intervening indels[2,17], these DNA recording systems are limited by their information-encoding capacity and durability. In other words, existing DNA recording systems erase DNA as their primary mode of recording information; although patterns of erasures contain new information, the inherent contradiction in removing DNA to add information creates fundamental challenges in the continued development of existing designs.

Ideally, a recorder would be capable of accumulating arbitrarily large amounts of new information through a series of updates that do not change or reverse previous updates. This could be accomplished by a DNA writing system in which insertions are sequentially added to a locus without disrupting previous insertions. We present a first version of such a recorder in CHYRON (Fig. 1). CHYRON combines a Cas9 nuclease with an hgRNA and a template-independent DNA polymerase, terminal

[1]Department of Biomedical Engineering, University of California, Irvine, Irvine, CA, USA. [2]NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine, Irvine, CA, USA. [3]Department of Computer Science, University of California, Irvine, Irvine, CA, USA. [4]Department of Chemistry, University of California, Irvine, Irvine, CA, USA. [5]Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine, CA, USA. [6]Center for Complex Biological Systems, University of California, Irvine, Irvine, CA, USA. ✉e-mail: ccl@uci.edu

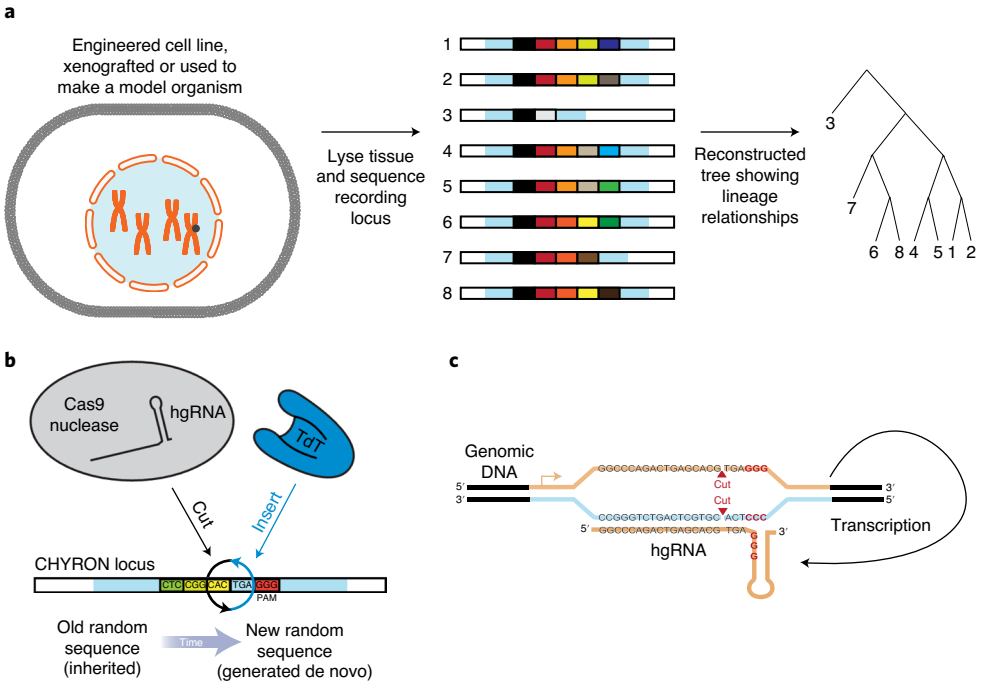


Fig. 1 | CHYRON: accumulation of insertion mutations in temporal order for the recording of cell history. The constitutive expression of Cas9 and TdT mediates the ordered acquisition of random insertion mutations, represented as differently colored boxes, written by TdT. **a**, Each cell is represented by one unique sequence, accumulated at the marked synthetic locus, which can be compared to other sequences to reconstruct the lineage of the cells. **b**, The CHYRON locus is cut by Cas9, insertions are added by TdT, and the process repeats to generate a series of ordered insertions. **c**, The CHYRON locus encodes an hgRNA, which directs Cas9 to the DNA that encodes it[3,4].

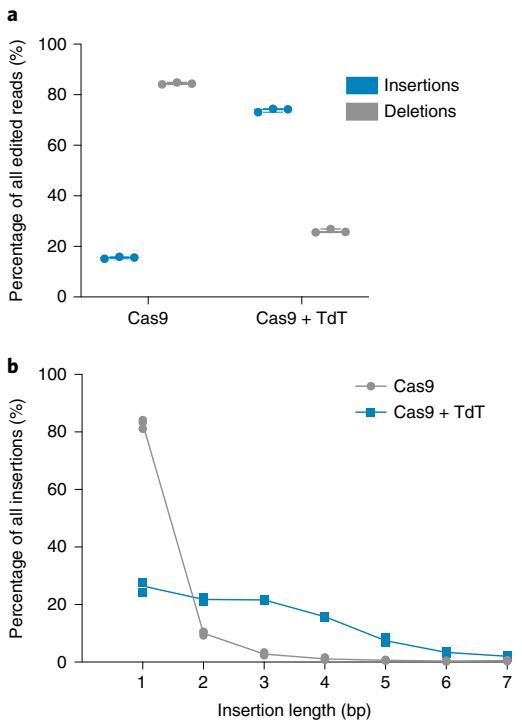
deoxynucleotidyl transferase (TdT)[19,20], which we show can efficiently add random nucleotides at Cas9-induced double-strand breaks (DSBs) (Fig. 2). The newly added nucleotides are then incorporated into the repaired DSB to produce a heritable insertion mutation consisting of random base pairs. Because an hgRNA repeatedly directs Cas9 to cut the locus encoding the hgRNA at a defined location relative to the protospacer adjacent motif (PAM)[3,4], cycles of cutting, nucleotide insertion by TdT and repair result in progressive and ordered insertional mutagenesis (Fig. 1 and Extended Data Fig. 1). We describe the successful implementation of CHYRON and its application to lineage reconstruction and the recording of hypoxia. We find that the information generated at a single CHYRON locus of <100 base pairs (bp) is sufficient to reconstruct the relatedness of populations containing hundreds of lineages as well as to report on the duration of exposure to a hypoxia mimic. This opens up the possibility of following development and profiling heterogeneous responses of cells to unevenly distributed or dynamic stresses at high cellular resolution in animals.

Results

TdT promotes random insertion mutations at Cas9 cuts. The core functionality of CHYRON relies on insertional mutagenesis mediated by TdT. We first tested whether a single round of Cas9 cutting and DSB repair could be intercepted by TdT to generate insertions. In HEK293T cells, we targeted Cas9 to a genomic locus in the presence or absence of TdT, expressed as a separate polypeptide. 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 (Methods). We found that, without TdT, the dominant mutation present at the Cas9 target site was a deletion (84%) or a 1-bp insertion (13%) (Fig. 2a,b), consistent with previous literature[21]. With TdT, the dominant mutations were insertions (74%) (Fig. 2a) with an average length of 2–4 bp (Fig. 2b and Supplementary Table 1); this pattern was conserved when the same genomic site was targeted in primary human fibroblasts (Extended Data Fig. 2a,b) and across all Cas9-targeted genomic sites tested in HEK293T cells (Supplementary Fig. 1b and Supplementary Table 1). TdT promoted insertions most efficiently when expressed without fusion to Cas9 or any other protein (Supplementary Fig. 2c). The dramatic effect of TdT on repair outcomes in a single round of editing suggested to us that, once we added an hgRNA, CHYRON would be able to write new DNA over multiple rounds with only moderate sequence corruption or loss through deletions.

TdT-mediated insertions must be random for CHYRON to mark individual cells distinctly during recording. We characterized the bias in base pairs inserted at various genomic target sites in the presence of Cas9 and TdT. The sites were chosen to represent all 16 possible pairs of −4 and −3 nucleotides relative to the PAM, as Cas9 canonically creates a blunt cut between these −4 and −3 positions and there is a known influence of these flanking nucleotides on editing outcomes[11,22–24]. We determined the proportions of each base pair in insertion mutations produced at these sites in the presence of TdT and calculated a Shannon entropy[25] of 1.9 bits per bp (Fig. 2c and Supplementary Table 1). This value is lower than the 2 bits of information that a single DNA base pair maximally contains, reflecting TdT's known bias for adding G nucleotides[26]. However, this deviation from perfectly random base pair insertion is slight. When we tested TdT insertion at the same site in HEK293T cells and primary dermal fibroblasts, we observed nearly identical distributions of inserted nucleotides (Extended Data Fig. 2c). We conclude that TdT has the potential to generate insertions containing high levels of information at nearly maximal information density for DNA.

CHYRON$_{20}$ loci accumulate multiple insertions in order. A recording locus should autonomously accumulate mutations



c

	A	C	G	T
AA*	16	26	47	11
AC**	17	26	46	11
AG	24	43	23	11
AT	13	40	35	13
CA	16	31	38	15
CC	13	41	26	20
CG	14	41	30	15
CT*	8	49	17	26
GA	9	37	35	19
GC	17	40	28	15
GG	24	18	47	11
GT	7	53	18	22
TA	13	36	38	14
TC	17	31	36	16
TG*	14	39	28	19
TT*	18	35	34	13

Fig. 2 | TdT writes stretches of random nucleotides at a Cas9-induced DSB. a, Expression of TdT promoted insertion mutations. HEK293T cells were transfected with plasmids expressing Cas9 and TdT or Cas9 alone and an sgRNA against a genomic site (HEK293 site 3). Three days later, cells were collected, DNA was extracted, and the targeted genomic site was amplified by PCR and sequenced by NGS. Each point represents a biological replicate; each biological replicate was carried out with two technical replicates, the mean of which is plotted. **b**, Expression of TdT resulted in longer insertion mutations than those minority insertions created in the presence of Cas9 alone, suggesting that TdT acts as a DNA writer. Of the pool of pure insertions, the percentage of each length was calculated and plotted. Each point represents the mean of two technical replicates of a single biological replicate (some points overlap; three biological replicates were assayed). **c**, Insertion sequences generated by TdT were random but had a bias toward G and C nucleotides. HEK293T cells were transfected with plasmids expressing Cas9 and TdT, and one of 16 sgRNA species 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 (these nucleotide identities are shown on the left of each row). 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 those in the row marked with **, which are the average of two biological replicates.

over multiple rounds of activity so that cellular and developmental processes occurring over time can be captured. To achieve multiple rounds of DNA writing, we combined TdT with an hgRNA locus to establish CHYRON. Because Cas9-induced DSBs are consistently generated between the −4 and −3 nucleotides relative to the PAM of the hgRNA locus (Extended Data Fig. 3), rounds of TdT-mediated insertion mutations should follow in order when repeated (Extended Data Fig. 1). This makes CHYRON an ideal recording locus because new insertions will neither remove nor corrupt previous insertions and insertions will be directionally arranged in the exact order in which they are added, simplifying lineage inference and offering options for recording temporal patterns of stimuli (Discussion).

To demonstrate repeated and ordered insertional mutagenesis, we integrated an hgRNA locus, including a 20-nucleotide (nt) spacer, at a single site in HEK293T cells (Extended Data Fig. 4a). We call this locus $CHYRON_{20}$, using the subscript to distinguish this specific instantiation of CHYRON that uses a 20-nt hgRNA spacer from others discussed below. When the HEK293T cell line containing $CHYRON_{20}$ (293T-$CHYRON_{20}$) was transfected with a plasmid encoding Cas9 and TdT for 3 d, insertions accumulated at the locus over time (Fig. 3a,b). As expected, the $CHYRON_{20}$ locus encoded more information than a benchmark system[4] in which the same hgRNA was expressed in the presence of Cas9 alone (Extended Data Fig. 4c). To test whether multiple rounds of insertion were occurring, we compared our results using $CHYRON_{20}$ to those shown in Fig. 2, in which a genomic locus with the same spacer sequence as $CHYRON_{20}$ was targeted by a single-guide RNA (sgRNA) that allows for only one round of editing. We found fewer 1–2-bp insertions and more long insertions at the $CHYRON_{20}$ locus than those in the single-round locus (Fig. 3c), suggesting that the $CHYRON_{20}$ locus received multiple rounds of insertions. To prove multi-round insertion, we transfected 293T-$CHYRON_{20}$ cells with a plasmid encoding Cas9 and TdT, isolated clones that had gained an insertion and transfected them again to observe whether these clones could gain an additional insertion (Extended Data Fig. 5a). Although editing was inefficient, new insertions were abundantly observed, and new insertions were found precisely downstream of the original insertion (Fig. 3d, Extended Data Fig. 5b and Supplementary Fig. 3).

$CHYRON_{20}$ showed our basic desired behavior, progressively inserting short random base pair stretches in order, but it was clear that $CHYRON_{20}$ could be improved. In particular, we deduced that $CHYRON_{20}$ only underwent an average of ~1.5 rounds of insertion, because the average insertion length at the end of the experiment,

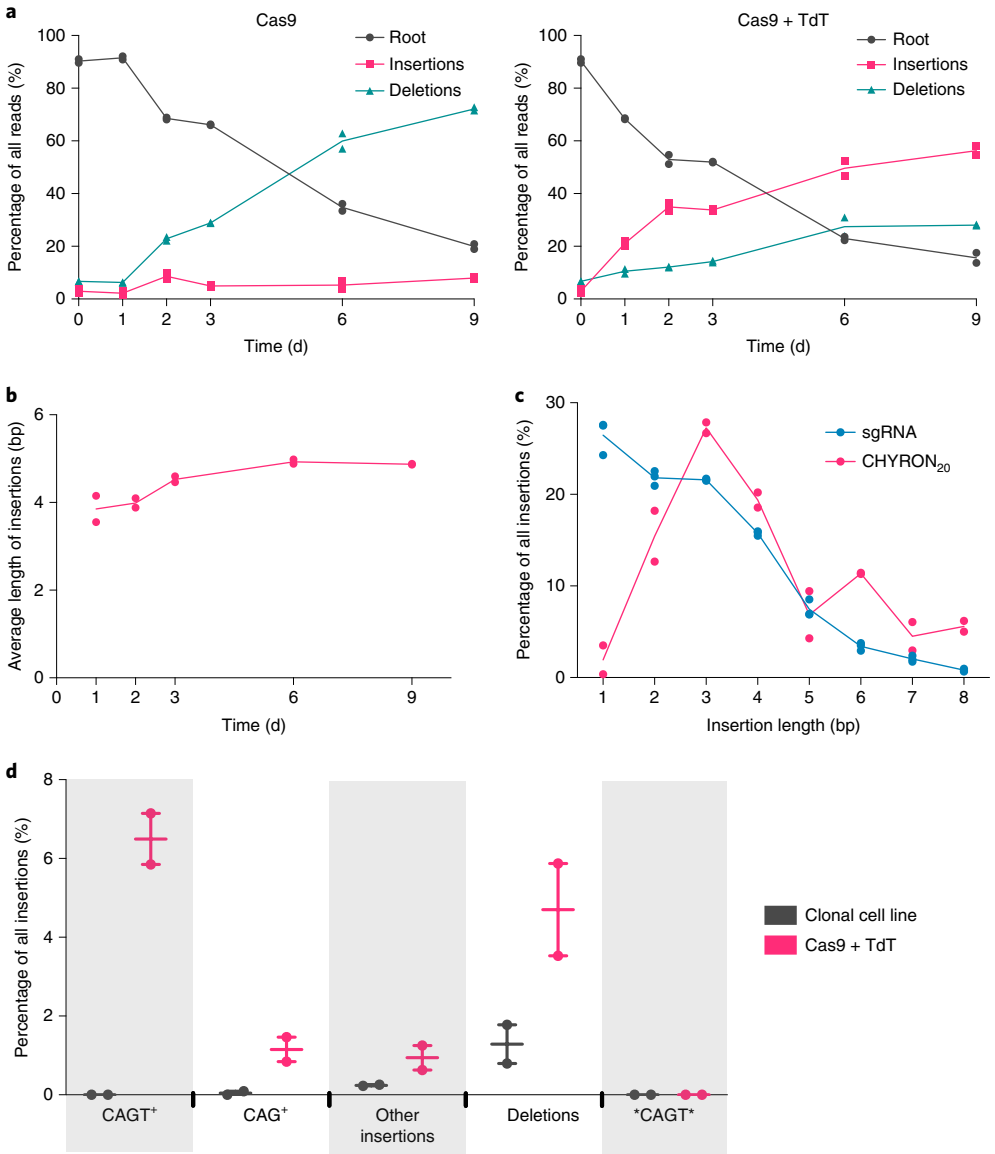


Fig. 3 | An integrated hgRNA accumulates insertions in multiple rounds. a, A clonal 293T-CHYRON$_{20}$ cell line bearing an integrated hgRNA was created so that expression of Cas9 and TdT resulted in multiple rounds of insertion of random nucleotides in an ordered fashion. The 293T-CHYRON$_{20}$ cells accumulated indels over time when exposed to Cas9 and TdT. The 293T-CHYRON$_{20}$ cells were transfected with a plasmid expressing Cas9 and, optionally, TdT for the indicated time before collection. Cells were retransfected every 3 d. The hgRNA locus was sequenced, and each sequence was classified into the following categories: unchanged (root), a pure insertion (insertions) or any sequence that involved a loss of information (deletions). Each point represents a single technical replicate (some points overlap; two replicates were assayed)**. b**, Insertions grew longer, on average, over time, until the 6-d time point, and then stopped growing. All sequences containing pure insertions were considered, and the average lengths were calculated. Each point represents a single technical replicate (some points overlap; two replicates were assayed). **c**, Longer insertions were more abundant for an hgRNA than for a protospacer targeted in a single round. Insertion lengths at the CHYRON$_{20}$ locus after 3 d of Cas9 and TdT expression were compared to insertion lengths at a genomic site with the same spacer sequence targeted with an sgRNA (data from Fig. 2b). Each point represents a single technical replicate (some points overlap; two replicates were assayed). **d**, Cas9 and TdT mediated multiple rounds of editing on an integrated hgRNA. The 293T-CHYRON$_{20}$ cell line was transfected to express Cas9 and TdT to induce insertions, and then a single colony was isolated. This new 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 15 d and then collected. The CHYRON locus was sequenced, and editing outcomes were determined to be the root CHYRON$_{20}$ sequence (not shown), deletions, the dominant CAGT insertion (not shown), 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. Each point represents a single biological replicate.

considering all sequences that had insertions, was 4.9 bp (Fig. 3b), whereas a single round of insertions generated only 2.88 bp on average (Supplementary Table 1). We next sought to improve CHYRON to be capable of more rounds of activity to enable a greater recording capacity.

CHYRON$_{16i}$ accumulates an average of 8.4 inserted base pairs. The failure of 293T-CHYRON$_{20}$ cells to write more than an average of 4.9 bp (Fig. 3b) has two likely explanations: (1) reduced efficiency of longer hgRNA species that are the product of rounds of TdT-mediated insertions and (2) silencing of the CHYRON locus.

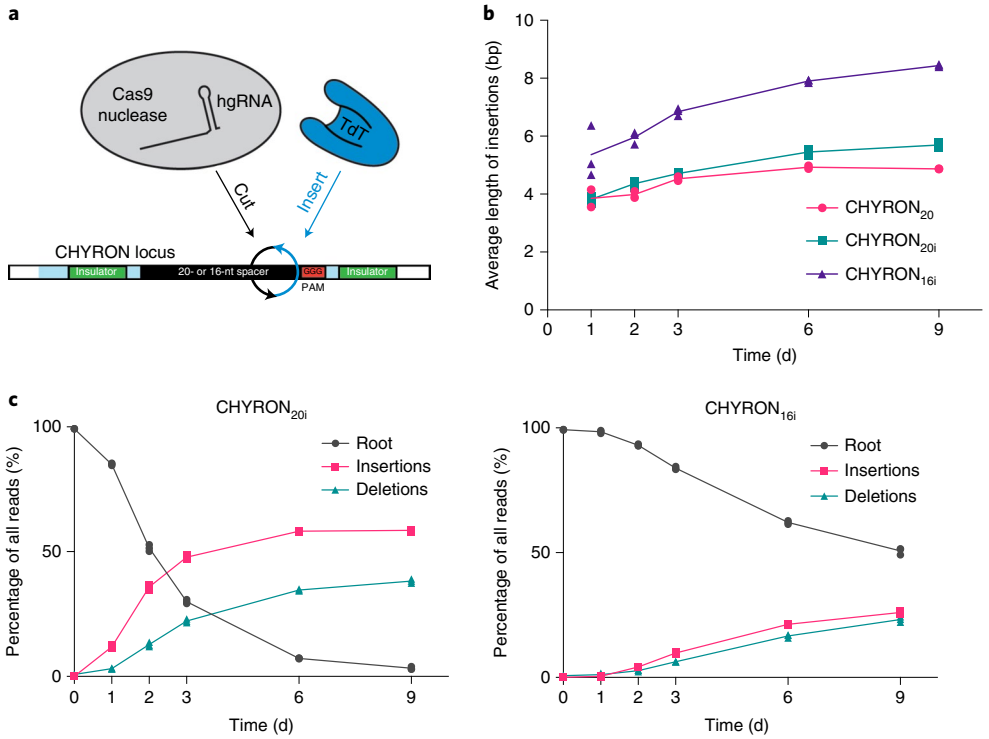


Fig. 4 | CHYRON writes an average of 8.4 bp in multiple rounds. a, Clonal HEK293T cell lines (hereafter 293T-CHYRON$_{20i}$ and 293T-CHYRON$_{16i}$) were created by integrating cassettes at the *AAVS1* safe-harbor locus. The cassettes contain hgRNA species with initial lengths of 20 or 16 nt, flanked by insulator sequences to prevent silencing. **b**, The CHYRON architecture allowed multiple rounds of cutting by Cas9 and writing by TdT. The 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 retransfected every 3 d. The CHYRON locus was analyzed by NGS, and the average lengths of insertions were calculated, considering all insertion-containing sequences. For CHYRON$_{20i}$ and CHYRON$_{16i}$, each point represents a single technical replicate (some points overlap; three replicates were assayed). Data from Fig. 3b is also plotted for reference. **c**, CHYRON loci with an initial hgRNA length of 20 nt accumulated indels over 6 d, whereas those with an initial hgRNA length of 16 nt accumulated indels more slowly but continued to do so over a 9-d time course. From the experiment described in **b**, each sequence was classified into the following categories: root, pure insertion (insertions) or any sequence that involved a loss of information (deletions). Each point represents a single technical replicate (some points overlap; three replicates were assayed).

To address these potential problems, we created two new cell lines (Fig. 4a), 293T-CHYRON$_{20i}$ and 293T-CHYRON$_{16i}$. CHYRON$_{20i}$ starts with a 20-nt spacer, whereas CHYRON$_{16i}$ starts with a 16-nt spacer. Both cell lines have the CHYRON locus flanked by chromatin insulator sequences[27] and integrated at the *AAVS1* safe-harbor locus in HEK293T cells. The 293T-CHYRON$_{16i}$ cells accumulated long insertions, reaching an average length for all insertion-containing sequences of 8.4 inserted base pairs, compared to 5.7 inserted base pairs for CHYRON$_{20i}$ (Fig. 4b). Note that 8.4 bp corresponds to a Shannon entropy of 14.6 bits (Extended Data Fig. 6a). As expected, CHYRON$_{16i}$ had a lower overall editing efficiency than did CHYRON$_{20i}$ (Fig. 4c), due to the low starting activity of shorter sgRNA species. Notably, CHYRON$_{16i}$ accumulated few short insertions (3–6 bp) compared to CHYRON$_{20i}$ (Extended Data Fig. 6b), suggesting that inhibition of Cas9 activity as the hgRNA lengthened indeed limited the duration of CHYRON activity in earlier designs. Silencing of the CHYRON locus was also likely important, because 293T-CHYRON$_{20i}$ cells continued to accumulate insertions throughout an entire 9-d time course (Fig. 4c). This is in contrast to CHYRON$_{20}$, which plateaued by 6 d and reached a final length of only 4.9 inserted base pairs, suggesting the utility of insulator sequences and the safe-harbor locus for CHYRON$_{20i}$. Consistently, CHYRON$_{17}$, which was delivered by lentivirus but not otherwise insulated, started with a 17-nt spacer and plateaued at an average length of 5.14 inserted base pairs in primary dermal fibroblasts (Extended Data Fig. 7c). These results suggest that both starting spacer length and expression context affect CHYRON performance.

CHYRON allows reconstruction of cell population lineages. To mimic a process of growth and spatial expansion over several days in a setting where we could know the ground-truth lineage relationships among populations of cells, we (1) grew ~40,000 293T-CHYRON$_{16i}$ cells bearing the root CHYRON$_{16i}$ sequence in four separate wells, (2) expressed Cas9 and TdT for 3 d (approximately three doublings) to allow the cells to write new base pairs at the CHYRON$_{16i}$ locus, (3) split each well into two and (4) repeated steps 2 and 3 again to yield 16 final wells (Fig. 5a). Cells in these final wells were allowed to grow for 3 d. By counting cells in an identical experiment on a hemocytometer, we estimated that there were 400,000 cells seeded in each of the final wells, which grew to approximately 3.2×10^6 cells per well by the time of collection. We subjected all cells from each final well to NGS of the CHYRON locus and found an average of 762 unique insertions per well (Supplementary Table 2). Using this NGS dataset, we inferred population lineage relationships by counting the number of shared sequences between pairs of wells to calculate relatedness (specifically, Jaccard similarity) and applying a standard agglomerative hierarchical clustering method to generate a tree from pairwise similarities (Fig. 5b). This resulted in the accurate and robust reconstruction of the full splitting procedure, validating CHYRON's application as an effective lineage tracer.

A practical consideration for any lineage-tracing locus is its performance under sampling constraints. In real settings, it is never possible to sequence from all cells in a relevant population due to inefficiencies in dissection, DNA extraction and amplification of the tracing locus. In vitro demonstrations of lineage tracing

typically bypass this issue by expanding cells to counter sampling inefficiency, but such post-experimental expansion is not possible for real biological samples such as terminally differentiated cells. Sampling inefficiencies are further compounded by the fact that, as we sample fewer and fewer molecules of any mutation-based lineage-tracing locus in a population, we expect the number of identical mutants arising independently in two populations (homoplasies) to decrease more slowly than the number of identical mutants that genuinely reflect the relatedness of two populations (Supplementary Note and Supplementary Fig. 4). Therefore, sampling inefficiency disproportionately reduces our ability to reconstruct population lineages from recording loci. CHYRON ameliorates this problem by achieving high levels of information through the long insertions generated, which reduces the chance of homoplasy. To test this, we computationally removed, at random, up to 75% of the unique insertion sequences in each well of our lineage-tracing experiment. Over ten trials in which different sequences were removed, we were still able to achieve near-perfect lineage reconstruction (Supplementary Fig. 4e). CHYRON should therefore have unique advantages in cases in which large populations and low sampling efficiency are a practical reality.

To test the limits of CHYRON lineage reconstruction more directly, we performed an additional population lineage-tracing experiment in culture (Extended Data Fig. 8a), using human adult primary dermal fibroblasts transduced with lentiviruses expressing the $CHYRON_{17}$ machinery, as shown in Extended Data Fig. 7a. The experimental plan was similar to that shown in Fig. 5, but the experiment began with 22,000 cells in each of four wells and ended with 190,000 cells in each of 16 wells (Extended Data Fig. 8a). With fewer cells in each final population and fewer cell divisions between each splitting event, an even smaller number of cells were predicted to acquire a unique CHYRON barcode, divide and then have their daughter cells deposited in and successfully sampled from different daughter wells. Despite this increased challenge and the reduced information-encoding capacity of $CHYRON_{17}$ relative to that of $CHYRON_{16i}$, we were able to reconstruct ten of 12 splits correctly (Extended Data Fig. 8b and Supplementary Fig. 5a); this performance was maintained in each of ten trials in which 20% of unique insertions were removed at random (Supplementary Fig. 5a).

We wished to compare the lineage-tracing performance of $CHYRON_{16i}$ to previously described DNA recording systems that record at a single site. A straight comparison of information generated predicted that CHYRON should lead to more robust lineage tracing, as the Shannon entropy of a CHYRON locus was higher than that of an hgRNA with Cas9 alone (Extended Data Fig. 4c), which is currently the DNA recording system with the highest single-site information (Supplementary Note). For example, in Kalhor et al.[11], mice were created in which each cell bore ~60 integrated hgRNA species and constitutively expressed Cas9. For the hgRNA encoding the most information (Methods), Kalhor et al. found 434 distinct mutated sequences in 45 mice. To provide an upper-bound estimate of the information encoded by this evolving hgRNA, we considered only edited sequences and assumed that each mutated sequence arose exactly once in each mouse in which it appeared. This resulted in a Shannon entropy of 7.97 bits (Supplementary Table 3), far lower than the 14.6 bits that can be encoded by $CHYRON_{16i}$ insertions (Fig. 4b and Extended Data Fig. 6a). As already discussed, high levels of information are useful in situations entailing large populations and low sampling efficiency, such as in our population lineage-tracing experiment with $CHYRON_{16i}$ (Fig. 5). Therefore, CHYRON should perform better than an hgRNA system alone.

To directly test this assertion, we computationally reduced our lineage-tracing NGS dataset (Fig. 5) to generate a 'truncated' dataset with the same information-encoding capacity as that of the hgRNA with the highest information content from Kalhor et al. (Methods).

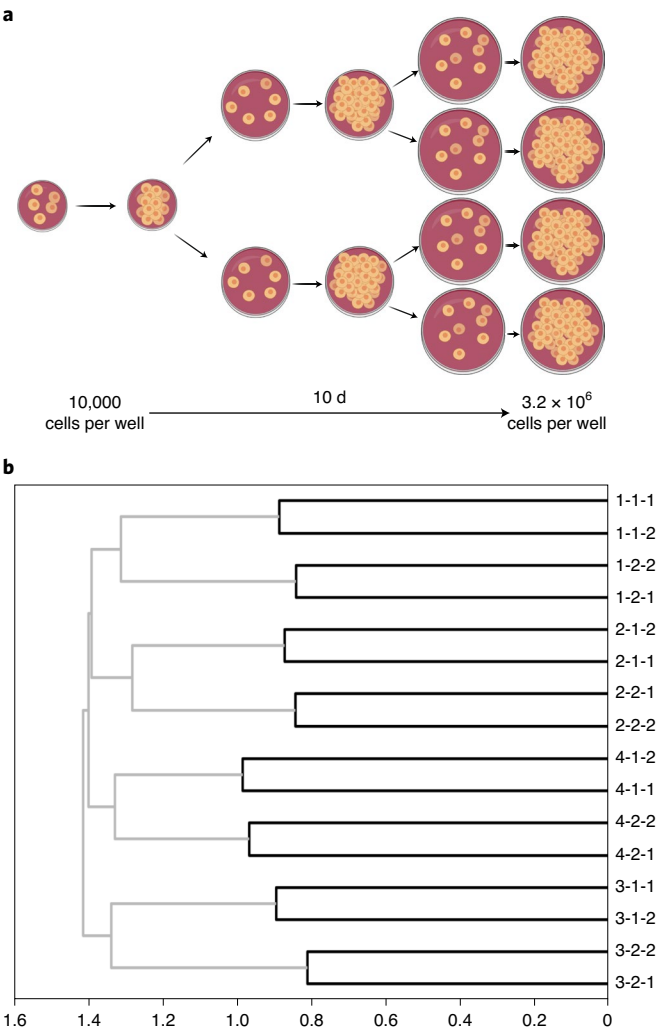


Fig. 5 | Reconstruction of cell relatedness by sequencing of the CHYRON locus. a, Schematic of the experiment. This experiment was performed in quadruplicate to yield 16 final wells. $CHYRON_{16i}$ cells were transfected to express Cas9 and TdT 1 d after plating, at days 1 and 5. **b**, A simple method led 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. Next, the Jaccard similarity coefficient between each pair of wells was computed, and hierarchical reconstruction was performed using the unweighted pair-group method with arithmetic mean (UPGMA) algorithm. Units are consistent across all plots in this work.

In essence, this truncated dataset recapitulates what would happen if we used the hgRNA system instead of CHYRON for the exact lineage-tracing experiment shown in Fig. 5. Although we found that both the CHYRON and truncated datasets were sufficient for perfect reconstruction of the relationships among the 16 populations when all the insertions from each well were used (Extended Data Fig. 9a), once we simulated more realistic sampling by computationally removing 80% of the insertion sequences from each well at random before truncation, the original CHYRON insertions resulted in a near-perfect reconstruction, whereas the truncated dataset resulted in highly inaccurate reconstructions (Extended Data Fig. 9). Therefore, we may conclude that, compared to other single-site DNA recorders, CHYRON is more effective for reconstructing population lineage relationships when sampling is limited, as it usually is in experimental contexts.

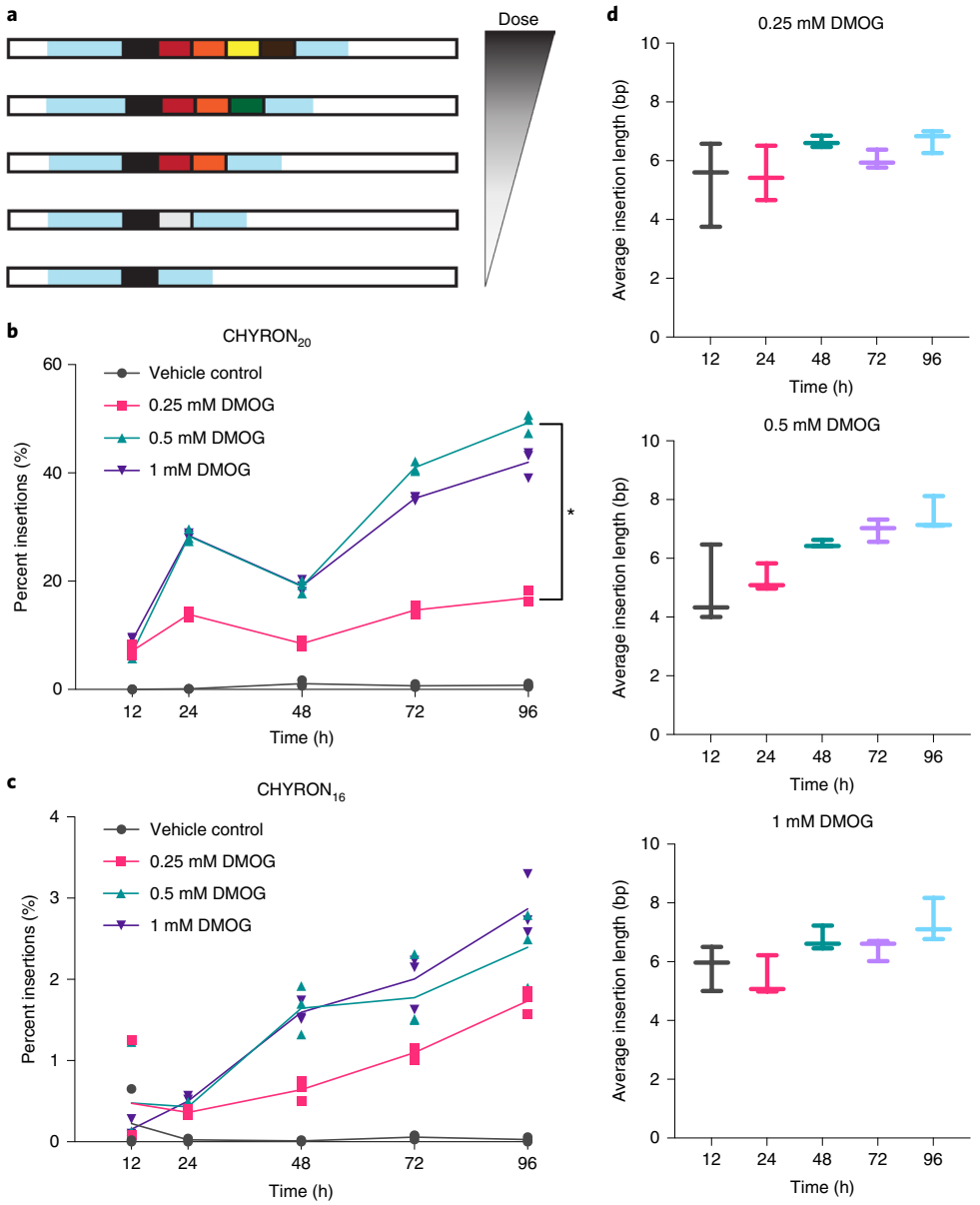


Fig. 6 | Expression of hypoxia-inducible Cas9 and TdT increases insertion abundance in proportion to dose and insertion length in proportion to duration of treatment with the hypoxia mimic DMOG. a, When recording is coupled to a cellular stress, the extent of insertions at the CHYRON locus can be used to reconstruct the extent of the stress. **b**, When transfected to express hypoxia-inducible Cas9 and TdT, $CHYRON_{20}$ loci accumulated insertions in a dose-dependent manner in response to DMOG. The 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 was additionally fused to a degron that destabilizes proteins in the presence of normal levels of oxygen. After transfection, cells were treated with DMOG or a vehicle control and then collected at the indicated time and analyzed as in Fig. 2a. Each point represents a single technical replicate (some points overlap; three replicates were assayed). For all time points except 12 h, the extent of insertions observed at the doses of 0.25 mM and 0.5 mM were significantly different by two-tailed *t*-test. For the 24-h time point, *t* value = 19.79, $P < 0.001$; for the 48-h time point, *t* value = 13.484, $P < 0.001$; for the 72-h time point, *t* value = 36.002, $P < 0.001$; for the 96-h time point, *t* value = 26.404, $P < 0.001$. **c**, By contrast, $CHYRON_{16}$ loci accumulated insertions at a lower, non-dose-dependent rate. The experiment was performed and analyzed, and data were plotted as in **b**. **d**, In 293T-$CHYRON_{16}$ cells transfected to express hypoxia-inducible Cas9 and TdT, insertions grew longer with increasing duration of exposure to DMOG. All sequences containing insertions were considered, and the lengths of the insertions were calculated. Each horizontal bar represents a single technical replicate.

CHYRON can report the dose or duration of hypoxia. Single-site DNA recording systems have been used to log cellular exposure to biological stimuli by making mutation at the recording locus inducible. For example, mutations at hgRNA species can be linked to inflammation exposure by placing Cas9 expression under the control of nuclear factor κB3 (ref. [3]). However, such hgRNA systems infer the dose or duration of the stimulus only at the population level, because it is difficult to determine the number of mutational cycles that each individual cell experiences when large deletions and deletions that terminate recording are common. Therefore, at the single-cell level, one can only reliably assess whether an hgRNA was mutated or not, providing dosage information only at the population level. While CHYRON can also carry out population-level stimulus recording, it offers the possibility of analog recording at the

level of a single cell, currently available only with substitution-based recorders that are less compact[8]. This is because the CHYRON locus progressively writes new base pairs and rarely erases, 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 (Fig. 6a).

To test stimulus recording with CHYRON, we linked insertional mutagenesis at a CHYRON locus to hypoxia, which triggers adaptive responses that affect a wide range of cellular behaviors[28,29]. We created a construct in which the expression of Cas9 and TdT is under the control of the 4× hypoxia-responsive element (HRE)-YB-TATA promoter[30], and in which Cas9 is additionally fused to an oxygen-dependent degron (ODD) domain[28]. Because these experiments were performed in parallel with those described in Fig. 4, we used the original 293T-CHYRON$_{20}$ cell line, described in Fig. 3, and the 293T-CHYRON$_{16}$ cell line, in which the CHYRON locus is integrated at *AAVS1* but no insulators are included. We transfected 293T-CHYRON$_{20}$ and 293T-CHYRON$_{16}$ cells with the 4×HRE-YB-TATA-Cas9-ODD-T2A-TdT construct and then exposed them to three different concentrations of the hypoxia mimic dimethyloxalylglycine (DMOG) for five different durations. In both cell lines, DMOG addition promoted insertions at the CHYRON locus (Fig. 6b,c). The proportion of the population bearing insertions increased with the dose of DMOG in 293T-CHYRON$_{20}$ cells (Fig. 6b). We note that the dose effect on the proportion of the 293T-CHYRON$_{16}$ population bearing insertions was not significant (Fig. 6c), likely because the overall low rate of insertions in this cell line reduced the dynamic range. Similarly, the length of insertions at the CHYRON$_{16}$ locus did not depend on the dose of DMOG (Fig. 6d). While initially surprising, this observation is consistent with a model in which dose determines the probability that CHYRON will become active in a cell but does not determine the degree of CHYRON activation. In keeping with this model, the average length of insertions at the CHYRON$_{16}$ locus increased significantly with duration of DMOG exposure (Fig. 6d and Extended Data Fig. 10). In other words, CHYRON was capable of recording exposure to stimuli in a manner that was digital (Fig. 6b) or analog (Fig. 6d), in which 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 digital and analog recording achieved with CHYRON is narrow (Fig. 6b–d and Extended Data Fig. 10), but, with further development, we expect that CHYRON will be an ideal system for capturing detailed cellular histories at single-cell resolution.

Discussion

There are two unique features of the CHYRON architecture that we believe will lead to its broad application and motivate its continued development in our and other laboratories. 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 in which it is difficult to capture all cells that might be related to each other; for these applications, a DNA recorder with a high information content is necessary to limit the possibility of misleading homoplasic 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. Using sequences of DNA recording loci to computationally reconstruct cell lineage remains challenging[31]; the ordered nature of CHYRON insertions may reduce the complexity of reconstructing cell lineage by clarifying when two cells diverged. Ordered insertions also make it possible that, if TdT can be engineered to add different types of nucleotides and the activity of the different TdTs can be coupled to different cellular stresses or the cell cycle, the CHYRON locus could record the relative timings of the different stresses in the cell's history or even provide an accurate count of cell divisions. The combination of these characteristics may enable single-cell-resolution lineage reconstruction from sparsely sampled CHYRON sequences, a goal that we are actively pursuing.

Combining CHYRON with other DNA recording innovations will yield substantial improvements in the depth and length of lineage recording. For example, if multiple CHYRON loci with different initial hgRNA spacer lengths were expressed in the same cell[11], CHYRON$_{20i}$ loci would fire early, while CHYRON$_{16i}$ loci would stochastically acquire an initial insertion that would increase the activity of the insertion-acquiring locus so that it would then accumulate additional insertions more quickly. Adapting CHYRON so that it can be read out by single-cell RNA sequencing[12,13,16] would allow information from multiple CHYRON loci to be combined to report on the lineage or history of a single cell. When encoded in an optimized genomic context, intermediate initial hgRNA lengths, such as 17 nt[32], are likely to pair higher initial activity than CHYRON$_{16i}$ with a greater recording capacity than that of CHYRON$_{20i}$, and thus could be a valuable tool as single- or multi-site recorders. These straightforward improvements to CHYRON could enable high-content-information and long-term recording in an organism (see Supplementary Note for additional considerations relevant to the use of CHYRON in vivo).

In the longer term, increasing the amount of information that CHYRON can encode and the durability of that information once encoded will allow the full potential of CHYRON to be realized. The information-encoding capacity, 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. This problem could be addressed by engineering Cas9 to better tolerate GC-rich sequences in its seed region or by using a different nuclease that cuts farther from its seed region. Additionally, the ~25% rate of deletion per Cas9 cut will eventually lead to information loss and inactivation of all CHYRON loci in the limit of prolonged continuous recording. Recruitment of factors that manipulate the balance of DSB repair pathways at the Cas9 cut site could reduce deletions substantially. The future development of CHYRON will be enhanced by wide interest in engineering new capabilities into its protein components, a CRISPR nuclease and TdT, in which there has been considerable recent interest as a tool for in vitro DNA synthesis[33–35]. Techniques that use polymerases[36–38], including TdT[39], to record time-series information on DNA synthesis timescales in vitro could also be merged with CHYRON. We predict that the unique components of CHYRON and the promise of the CHYRON architecture for reaching fully continuous recording of biological stimuli or lineage relationships at single-cell resolution in vivo will spur its continued development and application.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41589-021-00769-8.

Received: 20 January 2020; Accepted: 9 February 2021;
Published online: 22 March 2021

References

1. McDole, K. et al. In toto imaging and reconstruction of post-implantation mouse development at the single-cell level. *Cell* **175**, 859–876 (2018).
2. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).

3. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR–Cas in human cells. *Science* **353**, aag0511 (2016).
4. Kalhor, R., Mali, P. & Church, G. M. Rapidly evolving homing CRISPR barcodes. *Nat. Methods* **14**, 195–200 (2017).
5. Frieda, K. L. et al. Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).
6. 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).
7. Sheth, R. U., Yim, S. S., Wu, F. L. & Wang, H. H. Multiplex recording of cellular events over time on CRISPR biological tape. *Science* **358**, 1457–1461 (2017).
8. Tang, W. & Liu, D. R. Rewritable multi-event analog recording in bacterial and mammalian cells. *Science* **360**, eaap8992 (2018).
9. 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).
10. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175 (2016).
11. Kalhor, R. et al. Developmental barcoding of whole mouse via homing CRISPR. *Science* **361**, eaat9804 (2018).
12. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
13. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
14. Sheth, R. U. & Wang, H. H. DNA-based memory devices for recording cellular events. *Nat. Rev. Genet.* **19**, 718–732 (2018).
15. Hwang, B. et al. Lineage tracing using a Cas9-deaminase barcoding system targeting endogenous L1 elements. *Nat. Commun.* **10**, 1234 (2019).
16. Chan, M. M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).
17. 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).
18. 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).
19. Landau, N. R., Schatz, D. G., Rosa, M. & Baltimore, D. Increased frequency of N-region insertion in a murine pre-B-cell line infected with a terminal deoxynucleotidyl transferase retroviral expression vector. *Mol. Cell Biol.* **7**, 3237–3243 (1987).
20. Pryor, J. M. et al. Ribonucleotide incorporation enables repair of chromosome breaks by nonhomologous end joining. *Science* **361**, 1126–1129 (2018).
21. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR–Cas9 system. *Science* **343**, 80–84 (2013).
22. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
23. Zuo, Z. & Liu, J. Cas9-catalyzed DNA cleavage generates staggered ends: evidence from molecular dynamics simulations. *Sci. Rep.* **6**, 37584 (2016).
24. Gisler, S. et al. Multiplexed Cas9 targeting reveals genomic location effects and gRNA-based staggered breaks influencing mutation efficiency. *Nat. Commun.* **10**, 1598 (2019).
25. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
26. Motea, E. A. & Berdis, A. J. Terminal deoxynucleotidyl transferase: the story of a misguided DNA polymerase. *Biochim. Biophys. Acta* **1804**, 1151–1166 (2010).
27. Liu, M. et al. Genomic discovery of potent chromatin insulators for human gene therapy. *Nat. Biotechnol.* **33**, 198–203 (2015).
28. Semenza, G. L. Hypoxia-inducible factors in physiology and medicine. *Cell* **148**, 399–408 (2012).
29. Rankin, E. B. & Giaccia, A. J. Hypoxic control of metastasis. *Science* **352**, 175–180 (2016).
30. 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).
31. McKenna, A. & Gagnon, J. A. Recording development with single cell dynamic lineage tracing. *Development* **146**, dev169730 (2019).
32. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR–Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32**, 279–284 (2014).
33. Palluk, S. et al. De novo DNA synthesis using polymerase–nucleotide conjugates. *Nat. Biotechnol.* **36**, 645–650 (2018).
34. 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* **11**, 102 (2020).
35. 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).
36. Zamft, B. M. et al. Measuring cation dependent DNA polymerase fidelity landscapes by deep sequencing. *PLoS ONE* **7**, e43876 (2012).
37. Marblestone, A. H. et al. Physical principles for scalable neural recording. *Front. Comput. Neurosci.* **7**, 137 (2013).
38. Glaser, J. I. et al. Statistical analysis of molecular signal recording. *PLoS Comput. Biol.* **9**, e1003145 (2013).
39. Bhan, N. J. et al. Recording temporal data onto DNA with minutes resolution. Preprint at *bioRxiv* https://doi.org/10.1101/634790 (2019).

Methods

Plasmid cloning. Cloning was performed by standard Gibson assembly, in vivo recombination and restriction–ligation cloning. All plasmids are listed in Supplementary Table 4 and are available along with full sequences at Addgene (http://www.addgene.org/browse/article/28203329/). All plasmids used for transfection were purified with the HP GenElute Midi or Mini kit (Sigma, NA0200 and NA0150).

PCR was performed with Q5 Hot Start High-Fidelity DNA Polymerase or Phusion Hot Start Flex DNA Polymerase (New England Biolabs (NEB)). All primers were purchased from Integrated DNA Technologies (IDT), and PCR reagents were provided by NEB.

The human codon-optimized *S. pyogenes* Cas9 DNA sequence was amplified by PCR from hCas9, which was a gift from G. Church (Harvard Medical School, Addgene plasmid 41815 (ref. [40])). The XTEN linker was cloned by PCR of pCMV-BE3, which was a gift from D. Liu (Harvard University, Addgene plasmid 73021 (ref. [41])). The *T2A* self-cleaving sequence was inserted through PCR by designing primers with overhangs containing the *T2A* sequence. The sequence encoding TdT was amplified from the cDNA of an acute lymphoblastic leukemia cell line, and this entire insert was cloned into a pcDNA3.1 backbone, yielding Cas9-XTEN-TdT or Cas9-T2A-TdT. Cas9-containing constructs with the pcDNA3.1 backbone were transformed into XL10-Gold Ultracompetent *Escherichia coli* (Agilent, 200315).

Hypoxia-inducible constructs were cloned by adding a 4×HRE-YB-TATA promoter to drive the expression of Cas9-T2A-TdT. The 4×HRE sequence was amplified by PCR from 4×HRE_v2_YB-TATA-Gluc-CMV_dsRed[30], which was a gift from Y. Chen, University of California, Los Angeles. In addition, the sequence for the ODD was cloned to fuse with that of the C terminus of Cas9. The ODD sequence was amplified from HA-HIF1alpha-wt-pBabe-puro, which was a gift from W. Kaelin (Harvard Medical School, Addgene plasmid 19365 (ref. [42])). 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 E. Campeau and P. Kaufman (University of Massachusetts Medical School, Addgene plasmid 17451 (ref. [43])).

To clone sgRNA plasmids, the spacer region of the desired sgRNA was inserted into the pSQT1313 expression plasmid, which was a gift from K. Joung (Massachusetts General Hospital, Addgene plasmid 53370 (ref. [44])), 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[3,4]. 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$_{16}$, CHYRON$_{20i}$ or CHYRON$_{16i}$) 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.

Cas9-TdT constructs containing different linkers were cloned by restriction enzyme digestion and ligation. All restriction enzymes and T4 DNA ligase were purchased from NEB. All vector digestions were treated with calf-intestinal alkaline phosphatase from NEB after complete digestion by the restriction enzymes. A base construct was cloned first by Gibson assembly to add restriction sites to the original Cas9-XTEN-TdT construct (NheI-SfiI-Cas9-KpnI-XTEN-SexAI-TdT0). The base construct was digested with KpnI and SexAI. Three separate PCRs were performed to yield a 5×Flag or a 5×GSA linker product with KpnI and SexAI restriction sites. The 5×Flag was present on a gBlock (IDT), and the 5×GSA linker (four repeats of the sequence GSAGSAAGSGEF and a final repeat with the sequence GSAGSAAGASGEGRP[45]) 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 Cas9-5×Flag-TdT or Cas9-5×GSA-TdT.

Additional PCR on the 5×Flag gBlock yielded 5×Flag flanked by KpnI sites, which was digested and then ligated into Cas9-5×Flag-TdT, resulting in Cas9-10×Flag-TdT. To clone Cas9-15×Flag-TdT, a subsequent PCR was performed on the gBlock to amplify the 5×Flag sequence with FseI restriction sites. The FseI restriction enzyme recognition sites were used to ligate the 5×Flag sequence into Cas9-10×Flag-TdT, yielding Cas9-15×Flag-TdT.

Control plasmids were cloned containing the sequence for catalytically dead TdT (dTdT). The DNA fragment encoding dTdT was prepared by introducing the D343E and D345E mutations[46,47] into the wild-type TdT sequence. Two control plasmids were cloned by Gibson assembly into the pcDNA3.1 backbone: Cas9-XTEN-dTdT and dTdT alone.

The pcDNA3.1-sfGFP construct was cloned by Gibson assembly of the superfolder GFP gene from the yeast toolkit[48] into pcDNA3.1.

Lentiviral plasmids were cloned by Gibson assembly and restriction–ligation in Stbl3 *E. coli* (Thermo Fisher Scientific). The lentiviral backbone for both constructs was from lentiCas9-EGFP, a gift from P. Sharp and F. Zhang (Massachusetts Institute of Technology and Broad Institute of MIT and Harvard, Addgene plasmid 63592 (ref. [49])). In lentivirus 1, the sequence for mNeonGreen, not used in this experiment, was cloned upstream of that for Cas9, separated by a P2A sequence, expressed under the control of the *EFS* promoter. The sequence encoding mNeonGreen was obtained from pmNeonGreenHO-G (Addgene plasmid 127912, a gift from I. Tanida, Juntendo University School of Medicine[50]). In lentivirus 2, an hgRNA locus was cloned upstream of the *EFS* promoter driving expression of Luc2-P2A-tdTomato-T2A-TdT. The sequence for Luc2-P2A-tdTomato was from pCDH-EF1-Luc2-P2A-tdTomato (Addgene plasmid 72486, a gift from K. Oka (Baylor College of Medicine, unpublished)). Neither Luc2 nor tdTomato expression was used in our experiments. pU6, the promoter driving expression of the hgRNA, was from pSQT1313 (Addgene plasmid 53370 (ref. [44]), a gift from K. Joung). The hgRNA sequence was followed by a 7T terminator, which was immediately followed by a random 12-bp sequence. This random sequence was not used in our experiments. Lentivirus was produced and titered by VectorBuilder.

Cell culture and transfection. HEK293T cells were obtained from ATCC (CRL-3216) but were not otherwise authenticated. Normal adult human primary dermal fibroblasts were obtained from ATCC (PCS-201-012, lot 70015617 was used for Extended Data Fig. 2, and lot 61683453 was used for Extended Data Figs. 7 and 8). All cells were cultured in DMEM, high glucose, GlutaMAX Supplement (Gibco, 10566024), supplemented 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 (ref. [41]). Nucleofection of primary dermal fibroblasts was performed with a Lonza 4D-Nucleofector, according to the manufacturer's instructions, using P2 solution with electroporation code DS-150. For each nucleofection, 100,000 cells in a 20-µl solution were mixed with 400 ng DNA, nucleofected in an X Kit S cuvette and then plated in one well of a 24-well plate.

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 Cas9^{EQR}, a gift from K. Joung (Addgene, 65772 (ref. [51])), 50 ng of a plasmid expressing an sgRNA against a sequence at HEK293 site 3 or *AAVS1* that can be cut by Cas9^{EQR} and 1.5 µl FuGENE. HEK293T cells were transfected in a 24-well dish, transformants were selected with 1–2 µg ml^{-1} 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). Primary dermal fibroblasts were obtained directly from ATCC, where they were shown to be negative for mycoplasma, and used immediately.

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 or −3 position relative to the PAM. HEK293T cells in one well of a six-well dish were transfected with 0.4 µg of the specific sgRNA and constructs Cas9 (0.92 µg), Cas9-T2A-TdT (1.09 µg) or Cas9-5×Flag-TdT (1.1 µg). To normalize the total amount of DNA transfected, the Cas9 and Cas9-T2A-TdT transfections were supplemented with 0.12 µg and 0.01 µg pcDNA3.1-sfGFP, respectively. The cells were collected 3 d post-transfection, and DNA was extracted as detailed below.

Long-term editing on the CHYRON locus. The 293T-CHYRON$_{20}$, 293T-CHYRON$_{20i}$ or 293T-CHYRON$_{16i}$ cells were transfected in six-well dishes with 2 µg of the Cas9-T2A-TdT construct. Cell were grown for 1, 2, 3, 6 or 9 d after transfection. For the 6- and 9-d time points, 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, for the time course shown in Fig. 3, the same experiment was performed with a Cas9-T2A-STOP-TdT plasmid, which has a stop codon five amino acids into the TdT sequence. Samples from all time points were collected from single wells of a six-well plate (Falcon, 08-772-1B).

Two-step editing with Cas9 and TdT via isolation of single colonies. The 293T-CHYRON$_{20}$ cells were transfected with equal amounts of plasmids expressing Cas9–5×Flag–TdT and free TdT and then plated sparsely and grown until visible single colonies were 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 each grown in two wells of six-well plates. For each cell line, 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 clonal (>80% of reads from the untransfected sample were a single insertion sequence) and successfully sequenced. All samples were collected, and the CHYRON locus was sequenced via unique molecular identifier (UMI) incorporation and NGS.

Lineage reconstruction assay and analysis. For the reconstruction shown in Fig. 5, 5,000 293T-CHYRON$_{16i}$ cells were plated in each of four wells of a 384-well plate and then transfected the next day ('day 1') with a plasmid expressing Cas9 and TdT (pcDNA-Cas9-T2A-TdT). On day 2, cells were trypsinized, and the entire content of each well on the 384-well plate was moved to a 96-well plate. On day 4, when the cells had expanded to approximately 86,000 cells per well, each well was split into two wells of a 24-well plate, allowed to attach for 1 d and then transfected again. On day 8, when the cells in each well had expanded to approximately 800,000 cells, each well was split into two wells of a 6-well dish. On day 11, all wells were collected and analyzed by amplicon sequencing without UMI incorporation.

For our initial analysis, the researchers performing the analysis (M.W.S. and B.S.A.) were not told which well was which. We created a list of all insertion sequences in each well. Each insertion had an 'abundance', based on the number of NGS reads that included that exact insertion sequence, and a length, equal to the number of base pairs 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 8–15 bp. From this list of insertions, we created a binary vector for each well, the length of which 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 contained a 1 for a particular insertion if that insertion was present in the refined list for that well or a 0 if that insertion was absent. We used these vectors to calculate the Jaccard similarity between each pair of wells[4] and 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).

All analyses were carried out in Python. Scripts and detailed instructions are available at https://github.com/liusynevolab/CHYRON-lineage.

The lineage reconstruction shown in Extended Data Fig. 8 was performed as described above with the following modifications. A total of 38,000 actively growing, attached primary dermal fibroblasts in each of two wells of a 24-well plate were infected with each of the two lentiviruses expressing CHYRON$_{17}$: 570,000 viral particles of the lentivirus expressing Cas9 and 1,140,000 viral particles of the TdT/hgRNA lentivirus. Three days later (day 0), each well had 44,375 cells, which were split into two wells. On day 4, there were 65,000 cells in each well, and cells from each well were again split into two wells. On day 9, the cells in each well (of a 24-well plate) were transferred to one well of a six-well plate. On day 18, each well had 190,000 cells, which were split into two wells. On day 22, when each well had approximately 190,000 cells, cells were collected for DNA extraction. From the DNA extraction until the initial analysis was completed, the researcher (T.B.L.) was unaware of the identity of the wells. Because unedited reads were not removed during library preparation (see below), the abundance cutoff for insertions included in the analysis was adjusted to 0.0024% to account for the lower proportion of insertions (compared to unedited reads).

Hypoxia recording assay. The 293T-CHYRON$_{20}$ and 293T-CHYRON$_{16}$ cells were transfected in six-well dishes with 2 µg of the 4×HRE-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 was extracted at 24 or 48 h after DMOG addition. At 48 h, cells were replated and retransfected 14 h later. After transfection (14 h), DMOG was added at the indicated concentrations, and then the cells were grown for 24 h before collection of the 72-h time point and 48 h before collection of the 96-h time point.

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

All protein gels were 1.0-mm, 15-well, 4–12% Bis-Tris gels (Invitrogen, Thermo Scientific, NP0323), and electrophoresis was performed in an XCell SureLock Mini-Cell Electrophoresis System (Thermo Fisher Scientific, EI0001) with 1× 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 Immobilon-P (MilliporeSigma, 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 with blotting for TdT. The respective membrane was incubated with either Guide-it Cas9 Polyclonal Antibody (Clontech, 632607, 1:1,000 dilution), anti-TdT antibody (Abcam, ab14772, 1:1,000 dilution) or anti-actin antibody (Abcam, ab14128, 1:1,000 dilution) for 3 h at room temperature. Western blots were then incubated with horseradish peroxidase-fused secondary antibodies. Anti-rabbit IgG (Sigma-Aldrich, A0545, 1:10,000 dilution) was used to bind the primary antibodies for Cas9 and TdT, while anti-mouse IgG (R&D Systems, HAF007, 1:1,000 dilution) was used to bind the primary antibody for 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).

Deep sequencing library preparation of a genomic locus. The following protocol was used for Fig. 2 and Supplementary Figs. 1 and 2. Genomic DNA was isolated with the QIAamp DNA Mini kit (Qiagen, 51304) unless otherwise specified. An alternative protocol was developed for the extraction of genomic DNA from mammalian cells and was used for the experiments shown in Supplementary Fig. 2. 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% Triton X-100 and 200 µg µl^{-1} proteinase K. The lysis reaction was incubated at 65 °C for 10 min, and a 1:4 mixture of 7× 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 adaptors and a 5–7-nt sample-specific barcode (Supplementary Table 4). To ensure that the reporting of TdT-mediated insertions (which are GC rich, Fig. 2c) was not skewed by our library preparation, we tested the relative amplification of synthetic templates with varying proportions of GC nucleotides by two polymerases (Supplementary Fig. 1c). As Q5 polymerase produced a less skewed result, we used it for all library preparation PCR steps, except when indicated. 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 performed with 100 ng nucleic acid. 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 the 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. From the pooled clean product, 10 ng per individual sample was sent to Quintara Biosciences or Fornax Bio and run on an Illumina MiSeq.

At the sequencing vendor, the libraries were purified by binding to AMPure beads (0.9 volume beads:1 volume sample (hereafter 0.9:1)) 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 purified using an agarose gel, 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 Reagent kit version 2 (Illumina, MS-102-2003).

Deep sequencing library preparation of a genomic locus from primary cells. For Extended Data Fig. 2, genomic DNA was isolated with the QIAamp DNA Micro kit (Qiagen, 56304). After DNA extraction, the region targeted by Cas9 (site 3) was amplified by PCR. The primers contained the Illumina adaptors and a 5–7-nt sample-specific barcode (Supplementary Table 4). The initial PCR was performed for 30 cycles with Phusion Hot Start Flex polymerase in GC buffer (NEB), and then the technical replicates for each sample were pooled and purified with AMPure beads (0.9:1).

The libraries were sent to Genewiz 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-bp protocol.

Deep sequencing library preparation of the CHYRON locus at high efficiency for lineage reconstruction. For Fig. 5, genomic DNA was extracted with the QIAamp DNA Mini kit, and the entire recovered content was used in the initial PCR. The initial PCR was performed for 25 cycles with Phusion Hot Start Flex polymerase in GC buffer (NEB) and each sample was purified with AMPure beads (0.9:1) and then digested with PmlI (NEB) for 4 h and then purified with AMPure beads again. Next, the reamplification PCR was performed for 15–25 cycles with Q5 Hot Start polymerase. The samples were pooled according to their estimated concentration on an agarose gel stained with ethidium bromide and then purified with AMPure beads and cut with PmlI for an additional 4 h. Finally, bands of the expected library size or up to 100 bp larger were gel purified using a Macherey-Nagel PCR Clean-up kit. Purified products were sequenced on an Illumina HiSeq 2500 using the PE100 kit at the UCI Genomics High Throughput Facility.

Deep sequencing library preparation of the CHYRON locus expressed from a lentivirus in primary cells. For Extended Data Figs. 7 and 8, genomic DNA was

extracted with the QIAamp DNA Mini kit (for lineage reconstruction samples) or the QIAamp DNA Micro kit (for other samples). For lineage reconstruction samples, all recovered DNA was used in the initial PCR. The initial PCR was performed for 30 cycles with Phusion Hot Start Flex polymerase in GC buffer (NEB), and then each sample was purified with AMPure beads (0.9:1). The reamplification PCR was performed for 15 cycles with Q5 Hot Start polymerase (NEB). The samples were pooled according to their estimated concentration on an agarose gel stained with ethidium bromide and then purified with AMPure beads as described before. Purified products were sequenced on an Illumina HiSeq 4000 using the PE150 kit by Novogene.

Unique molecular identifier incorporation for sequencing of the CHYRON locus. For Figs. 3, 4 and 6, UMIs of 20 degenerate nucleotides were incorporated to barcode the initial primer extension product from each individual DNA template containing the integrated CHYRON locus. Primers were ordered from IDT containing the following: the Illumina reverse adaptor, a UMI, a 5–7-nt sample-specific barcode and an hgRNA construct-binding region (Supplementary Table 4). Genomic DNA was isolated with the QIAamp DNA Mini kit (Qiagen), and 600 ng nucleic acid was used. The UMI incorporation reaction was run with Phusion Hot Start Flex DNA Polymerase (NEB) under the following conditions: 98 °C, 5 min; (55 °C, 30 s at a ramp rate of 4 °C s^{-1}; 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 min at 37 °C.

A downstream PCR was performed after the UMI incorporation step to amplify specific sequences that contained a UMI. The sample was run with a forward primer with the Illumina forward adaptor, a 5–7-nt sample-specific barcode, and an hgRNA-binding region and a reverse primer complementary to the Illumina reverse adaptor 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 ramp rate of 2 °C s^{-1}. 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 described for genomic sites for Fig. 3b–d. For the rest of the experiments, samples were individually purified with AMPure beads (0.9:1) and then reamplified for 15 cycles, pooled and gel purified, including fragments ~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 Reagent kit version 2 (Illumina, MS-102-2003).

Amplification bias assay of varying polymerases. As 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 nt 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 adaptors. The PCRs 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 ramp rate of 2 °C s^{-1}. 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 Fornax Bio and sequenced as described above.

Deep sequencing analysis. The sequences retrieved by NGS were first grouped to individual samples based on their barcodes. Next, associated forward and reverse reads were merged for each sample (PEAR 0.9.10 (ref. [52])) and mapped to the reference sequence by the alignment algorithm implementation (Mapp) used in Perli et al.[3], which provides a sequence of M (match), X (mismatch), I (insertion) and D (deletion) as the mapping result.

If UMI barcodes were present, before mapping, sequences with the same barcodes were combined into one. To combine, we started with a multiple alignment of the sequences (carried out with Motility library in Python) with the same UMI barcode (a 1-nt 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, poor alignments (>20 mismatches or >50% deletions) were removed. Next, the positions of mismatches and inserted or deleted sequences on the reference sequence and their frequencies were extracted; sequences were placed in one of three categories: unchanged, pure insertions (insertions) or any sequence that led to a loss of information (deletions). 'Deletions' included pure deletions; mixtures of insertion, deletion and substitution mutations; and pure substitutions (2% of all edited sequences in the typical experiment shown in Fig. 2a). Only insertions or deletions occurring around (−10 nt to +10 nt) the cut site were kept (for the data in Fig. 2, Extended Data Fig. 2 and Supplementary Figs. 1 and 2, 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 (>12 nt) could map (with less than a 2-nt difference) to sequences of our plasmid backbones, they were filtered out. In addition, insertions longer than 15 nt (or 20 nt for Fig. 2c and 40 nt for Extended Data Fig. 6b) were excluded, as these were found to more frequently have nucleotide biases that suggested that they were TdT independent. For the data shown in Extended Data Figs. 2 and 7, substitution mutations were detected at a higher rate, likely due to our increased use of Phusion polymerase, which has a higher error rate than the usual Q5 polymerase, during library preparation. Therefore, the rate of substitution mutations observed in unedited negative control samples was subtracted from the rate observed in each edited sample.

All analyses were carried out in Python. The scripts are available at https://github.com/liusynevolab/CHYRON-NGS.

For the experiment shown in Fig. 5, because samples were sequenced with a paired-end 100-bp protocol, rather than the paired-end 150-bp or 250-bp protocol we used for all other experiments, forward and reverse reads were not paired, and only forward reads were used for the analysis.

Entropy calculations. To calculate Shannon entropy, we first made a table of all the unique sequences in the relevant dataset, with the number of times each sequence was observed (the 'count'). For each sample shown in Extended Data Fig. 4c, we extracted the sequence flanking the cut site (34 bp in the root sequence, shorter in loci with deletions, longer in those with insertions) from each read that aligned to the reference sequence. The number of unique UMIs associated with the sequence was considered as the count. For the calculations shown in Extended Data Fig. 6a, we used as an input all insertion sequences that appeared in at least 0.0139% of non-deletion reads in any well of our lineage-tracing dataset. Next, we used Python scripts to calculate how many times each possible 1-, 2-, 3- or 4-nt sequence appeared in these insertions. For example, if the insertion GAG appeared with 1,000 reads in our dataset, we would record a count of 1,000 for the sequences A, GA, AG and GAG and a count of 2,000 for the sequence G. Once a table of unique sequences and counts (c) was created, we calculated the proportion (p) for each sequence (equation shown here for sequence i).

$$p_i = \frac{c_i}{c_1 + \ldots + c_i \ldots + c_n} \quad (1)$$

Next, we calculated the overall Shannon entropy (H) for the dataset.

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (2)$$

All analyses were carried out in Python and Excel. The scripts are available at https://github.com/liusynevolab/CHYRON-entropycalc and https://github.com/liusynevolab/CHYRON-insertion-entropy-calc.

Simulation of lineage reconstruction with the information-encoding properties of an hgRNA with Cas9 alone. First, we identified the single hgRNA in Kalhor et al.[11] with the highest information-encoding capacity when mutated by Cas9 alone. To do this, we calculated the number of unique mutated sequences observed per mouse sequenced for each of the hgRNA species presented. Next, for the top four hgRNA species, we made a table in which we recorded the number of mice in which each unique mutated sequence was observed (which we considered as the count). We calculated the Shannon entropy from this table (as in Supplementary Table 3). We argue that the value we calculated constitutes the maximum Shannon entropy, because some common mutated sequences likely arose multiple times in the same mouse, but we only increased the count by one for each mouse in which the sequence appeared; decreasing proportions of more-common sequences lead to increased entropy. The hgRNA (36) with the highest Shannon entropy was used for the simulation. As described above, we considered the number of mice in which it appeared as the count for each mutated sequence. For each of the 434 mutated sequences that arose from hgRNA 36, we calculated the self-information (I) in bits.

$$I = p_i \log_2 p_i \quad (3)$$

Next, we calculated the length of a CHYRON insertion with the same self-information for each of the 434 sequences, using our earlier determination (Extended Data Fig. 6a) that CHYRON insertions encode 1.74 bits per bp. If the calculated length was not an integer, it was expressed as a proportion of integer lengths. For example, if the calculated self-information for sequence x is 4.35 bits, the equivalent insertion length is 4.35 bits × (1.74 bits per bp)$^{-1}$ = 2.5 bp. Thus, sequence x corresponds to 50% 2-bp insertions and 50% 3-bp insertions. The equivalent insertion lengths for all 434 mutated sequences were determined, and these values, weighted according to the abundance of the mutated sequence in the Kalhor et al. dataset, were used to calculate the proportions of insertions of each length required to match the overall information capacity and distribution in the hgRNA dataset. We calculated that a CHYRON dataset equivalent in information-encoding capacity to hgRNA 36 would include 3% 2-bp insertions, 19% 3-bp insertions, 17% 4-bp insertions, 40% 5-bp insertions and 21% 6-bp insertions.

For the next step in our simulation, we created a CHYRON dataset with the above insertion lengths. First, we created a 'complete list' of the insertions that were at least 6 bp in length and that represented at least 0.0139% of all non-deletion reads for each well in our lineage-tracing experiment (Fig. 5). Next, we created a 'dictionary table' of all unique insertions in this master list, along with all wells in which they were detected. At random, we truncated 3% of the insertions in the dictionary table to 2 bp, 19% to 3 bp, etc. All truncations removed the 'downstream' nucleotides. In this way, we created a 'truncated list' for each well. The truncated lists were used to calculate the Jaccard similarity between all possible pairs of wells and to perform UPGMA hierarchical clustering. The truncation process was repeated five times. For analyses in which sampling was limited, we started with the complete list for each well. From the complete list, we simulated poor sampling by discarding 80% of the insertions from each well at random to create the 20% list. Next, we proceeded to create the dictionary table and perform the truncations as described above to create the truncated 20% lists. In parallel, we used the (full-length) 20% lists and truncated 20% lists to calculate Jaccard similarities and perform UPGMA hierarchical reconstructions. The entire sampling, truncating and reconstruction processes were performed in five replicates.

All analyses were carried out in Python and Excel, and scripts are available at https://github.com/liusynevolab/CHYRON-truncation-simulation.

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 together, and DNA extractions were performed together. All procedures downstream of DNA extraction were performed separately. For the data shown in Fig. 6b and Extended Data Fig. 10, statistical analysis was performed in SPSS. For Fig. 6b, the values of the three technical replicates were listed for each sample. For comparing the 0.25-mM and 0.5-mM doses, the samples from each time point were compared by independent samples two-tailed *t*-tests (variances were equal as determined by Levene's test ($P > 0.05$)). For Extended Data Fig. 10, a list of all insertion lengths recorded was created for each time point and dose (for this analysis, technical replicates were pooled). Next, for each dose, we determined whether the lengths were significantly different between time points. For each dose, variances were unequal as determined by Levene's test ($P < 0.05$). Therefore, we performed Welch's one-way ANOVA, followed by a post hoc Games-Howell test.

Figure preparation. Fig. 5a and Extended Data Figs. 1, 5a, 7a and 8a were prepared at https://biorender.com/. The plots in Fig. 5b, Extended Data Figs. 8b and 9b and Supplementary Fig. 4a,c were generated using the 'hierarchy.dendrogram' function in matplotlib (https://scipy.org/). The threshold for color change was set at the default level: $0.7 \times \max(Z(:,2))$.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All NGS datasets were deposited at the NCBI's Sequence Read Archive under accession no. PRJNA561027. All plasmids and full sequences are available at Addgene. See Supplementary Table 4 for a guide to these data and reagents. Please contact C.C.L. for cell lines.

Code availability

All scripts are available at https://github.com/liusynevolab.

References

40. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
41. 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).
42. Yan, Q., Bartz, S., Mao, M., Li, L. & Kaelin, W. G. The hypoxia-inducible factor 2α N-terminal and C-terminal transactivation domains cooperate to promote renal tumorigenesis in vivo. *Mol. Cell Biol.* **27**, 2092–2102 (2007).
43. Campeau, E. et al. A versatile viral system for expression and depletion of proteins in mammalian cells. *PLoS ONE* **4**, e6529 (2009).
44. Tsai, S. Q. et al. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.* **32**, 569–576 (2014).
45. Waldo, G. S., Standish, B. M., Berendzen, J. & Terwilliger, T. C. Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* **17**, 691–695 (1999).
46. Yang, B., Gathy, K. N. & Coleman, M. S. Mutational analysis of residues in the nucleotide binding domain of human terminal deoxynucleotidyl transferase. *J. Biol. Chem.* **269**, 11859–11868 (1994).
47. Repasky, J. A. E., Corbett, E., Boboila, C. & Schatz, D. G. Mutational analysis of terminal deoxynucleotidyltransferase-mediated N-nucleotide addition in V(D)J recombination. *J. Immunol.* **172**, 5478–5488 (2004).
48. Lee, M. E., DeLoache, W. C., Cervantes, B. & Dueber, J. E. A highly characterized yeast toolkit for modular, multipart assembly. *ACS Synth. Biol.* **4**, 975–986 (2015).
49. Chen, S. et al. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell* **160**, 1246–1260 (2015).
50. Tanida-Miyake, E., Koike, M., Uchiyama, Y., Tanida, I. & Sato, M. Optimization of mNeonGreen for Homo sapiens increases its fluorescent intensity in mammalian cells. *PLoS ONE* **13**, e0191108 (2018).
51. Kleinstiver, B. P. et al. Engineered CRISPR–Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).
52. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2013).

Acknowledgements

We thank S.K. Paul and T.C. Lone for technical assistance. We thank the following people for helpful discussions: C. Guerrero-Juarez, C. Li, J. Zimak, Q. Nie and all members of the Liu laboratory. We thank the following people for plasmids: Y. Chen, K. Joung, W. Kaelin, G. Church, D. Liu, E. Campeau, P. Kaufman, T. Lu, P. Sharp, F. Zhang, K. Oka and I. Tanida. 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 C.C.L., AHA Predoctoral and NSF Graduate Research fellowships to C.K.C. and a fellowship from the NSF-Simons Center for Multiscale Cell Fate Research (NSF award 1763272) to T.B.L.

Author contributions

T.B.L. and C.C.L. designed experiments. C.K.C. performed NGS library preparation for hypoxia recording experiments; T.B.L. and J.H.G. performed all other experiments, with assistance from M.W.S. C.C.L. and T.B.L. developed hypoxia recording protocols, G.L. determined how to remove unedited CHYRON sequences during NGS library preparation, and J.H.G. and T.B.L. developed all other NGS library preparation protocols. T.B.L., J.H.G. and C.K.C. established cell lines. T.B.L., J.H.G., C.K.C., G.L. and M.F. cloned plasmid vectors. T.B.L., M.W.S., E.F., B.S.A., X.X. and C.C.L. discussed experimental analyses. M.W.S. and B.S.A. wrote lineage reconstruction scripts and then performed initial reconstructions for the experiment shown in Fig. 5; T.B.L. performed all other lineage analyses. E.F. wrote code for the analysis of NGS data, which was subsequently edited by M.W.S., B.L. and T.B.L. B.L., M.F. and T.B.L. analyzed the proportions of all 2-, 3- and 4-nt sequences in CHYRON insertions. E.F. wrote the description of NGS analysis in the Methods, and J.H.G. and T.B.L. wrote the remainder of the Methods. T.B.L. and C.C.L. wrote the remainder of the paper, with input from all authors, especially C.K.C. C.C.L. procured funding and oversaw the project.

Competing interests

The authors declare no competing interests.

Additional information

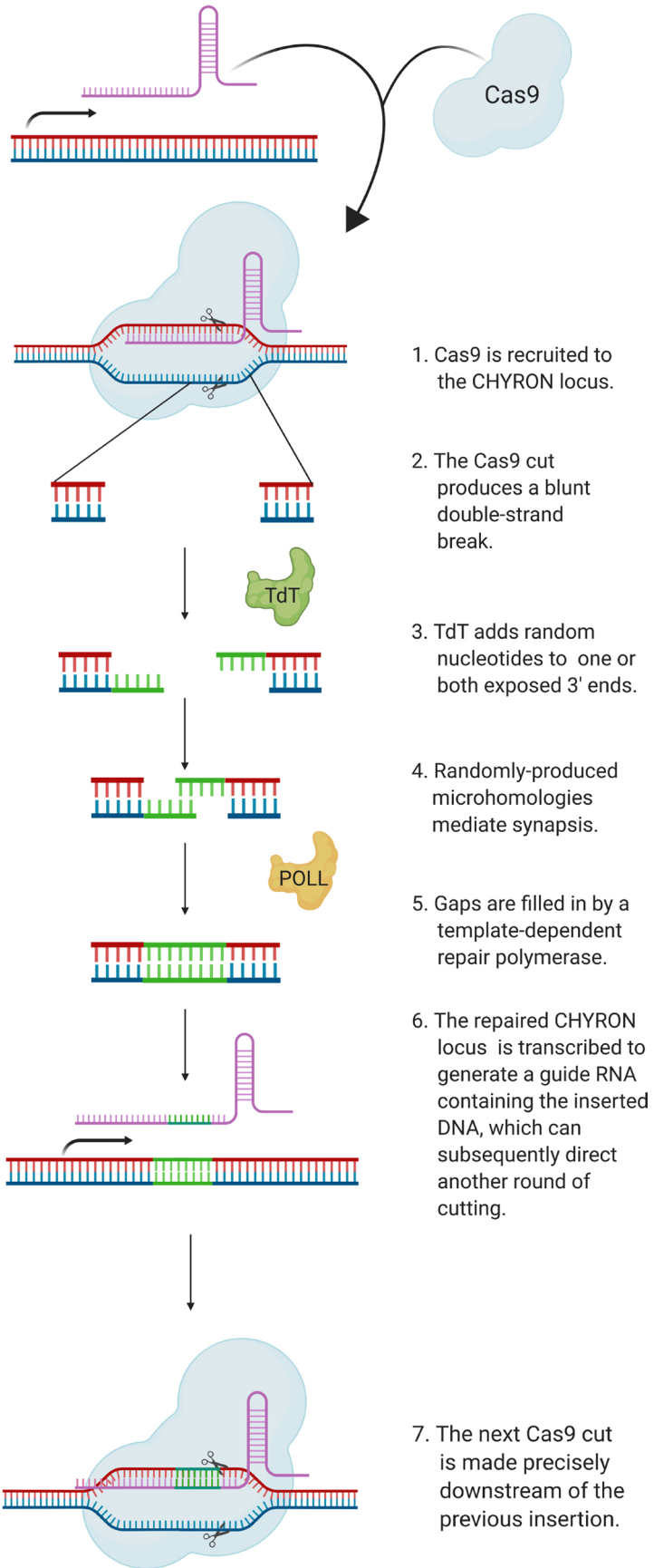
Extended data is available for this paper at https://doi.org/10.1038/s41589-021-00769-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41589-021-00769-8.

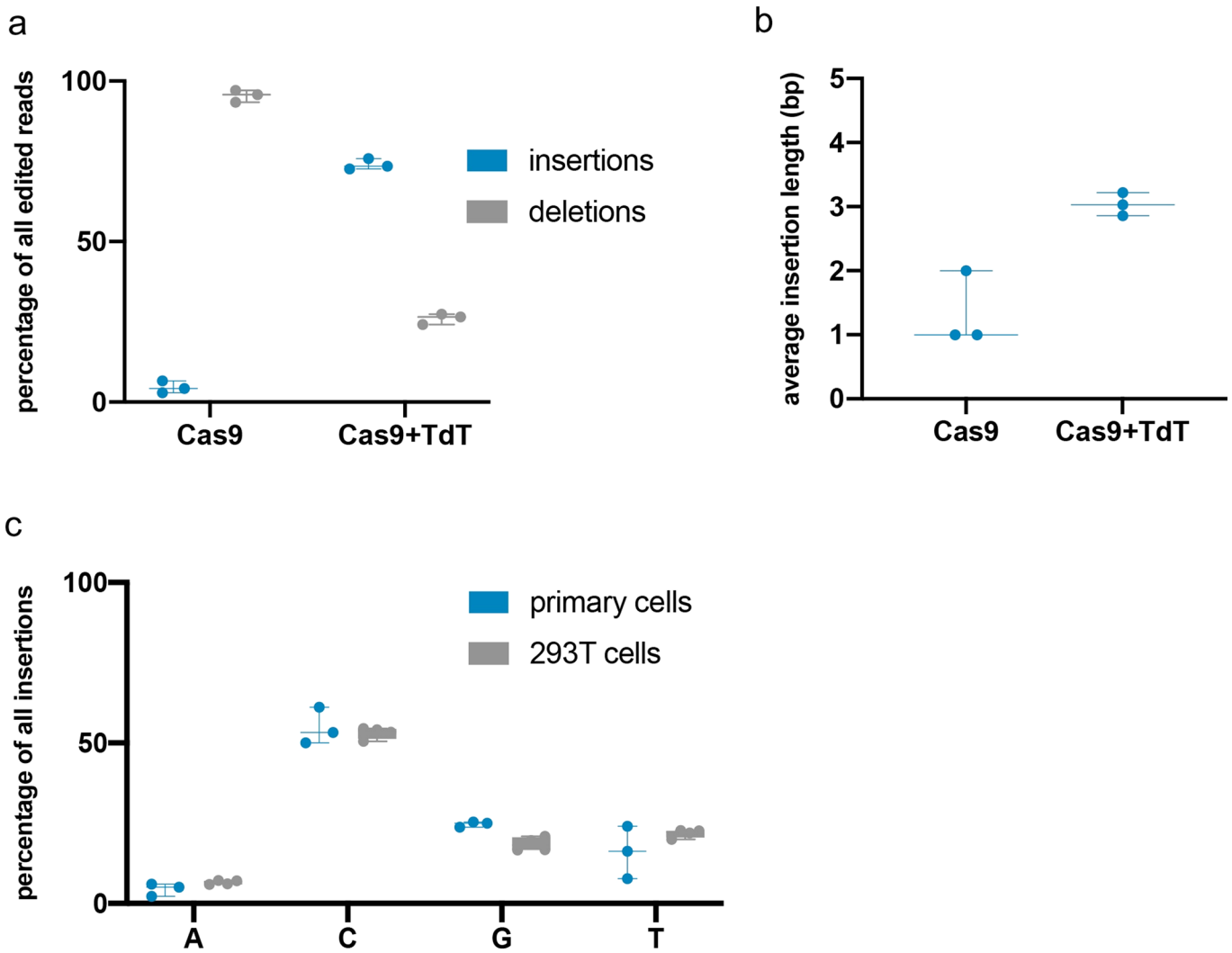
Correspondence and requests for materials should be addressed to C.C.L.

Peer review information *Nature Chemical Biology* thanks Randall Platt and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

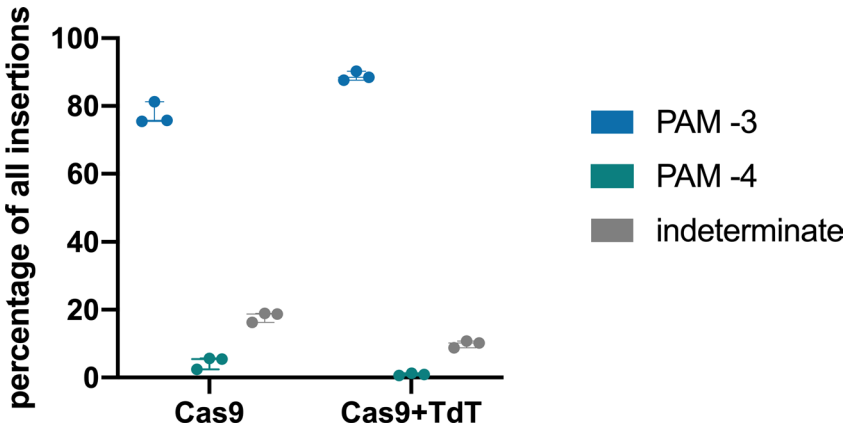
Reprints and permissions information is available at www.nature.com/reprints.



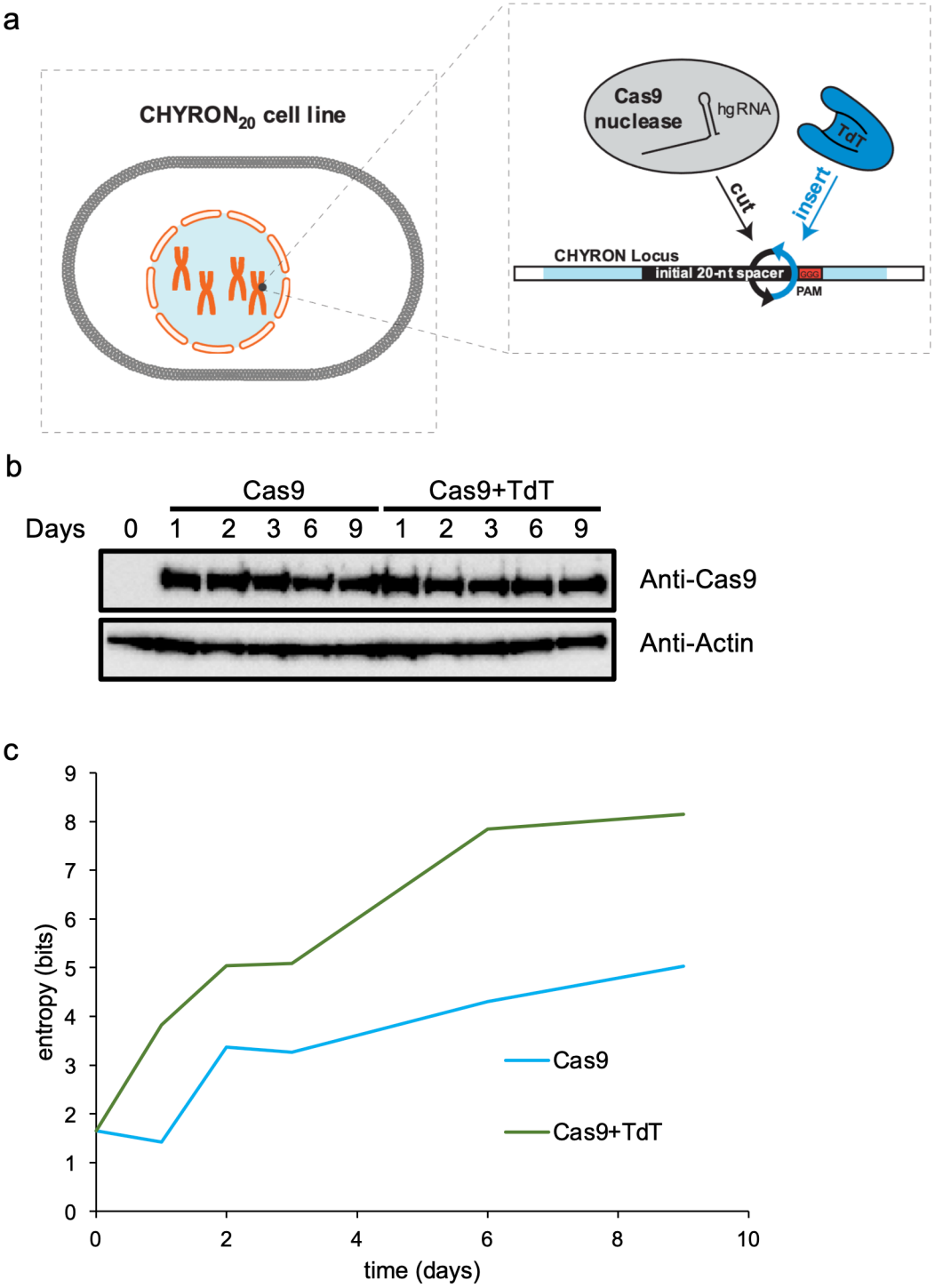
Extended Data Fig. 1 | Detailed model for the progressive accumulation of insertions at the CHYRON locus. The nucleotides initially added by TdT may be ribonucleotides[20].



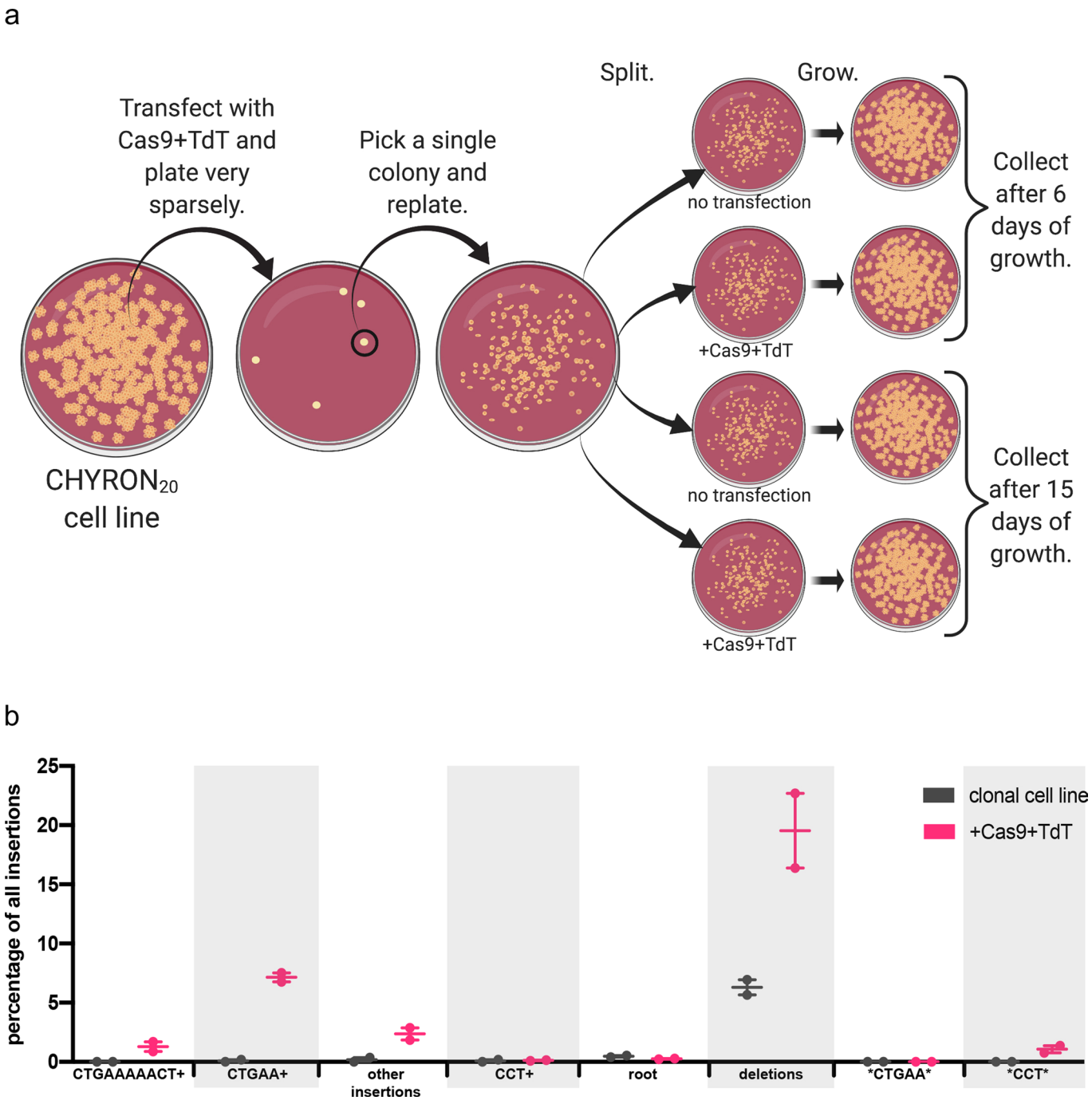
Extended Data Fig. 2 | TdT writes stretches of random nucleotides at a Cas9-induced DSB in primary cells. a, Expression of TdT promoted insertion mutations. Adult human primary dermal fibroblasts were nucleofected with plasmids expressing Cas9 and TdT, or Cas9 alone, and an sgRNA against a genomic site (site3, as in Fig. 2). After 7 days, cells were collected, processed, and analyzed as described in Fig. 2a and Methods. Each point represents a single technical replicate. (Three replicates were assayed.) **b**, Expression of TdT resulted in longer insertion mutations than those minority insertions created in the presence of Cas9 alone, suggesting that TdT acts as a DNA writer. From the pool of pure insertions, the average length was calculated and plotted. Each point represents a single technical replicate. (Three replicates were assayed.) **c**, Insertion sequences generated by TdT had the same nucleotide biases in primary fibroblasts as in HEK293T cells. The proportions of each nucleotide (on the top strand) found in all pure insertion sequences 4 bp in length were calculated and plotted. Each point represents a single technical replicate. (HEK293T data from Fig. 2c, sequence 'GT').



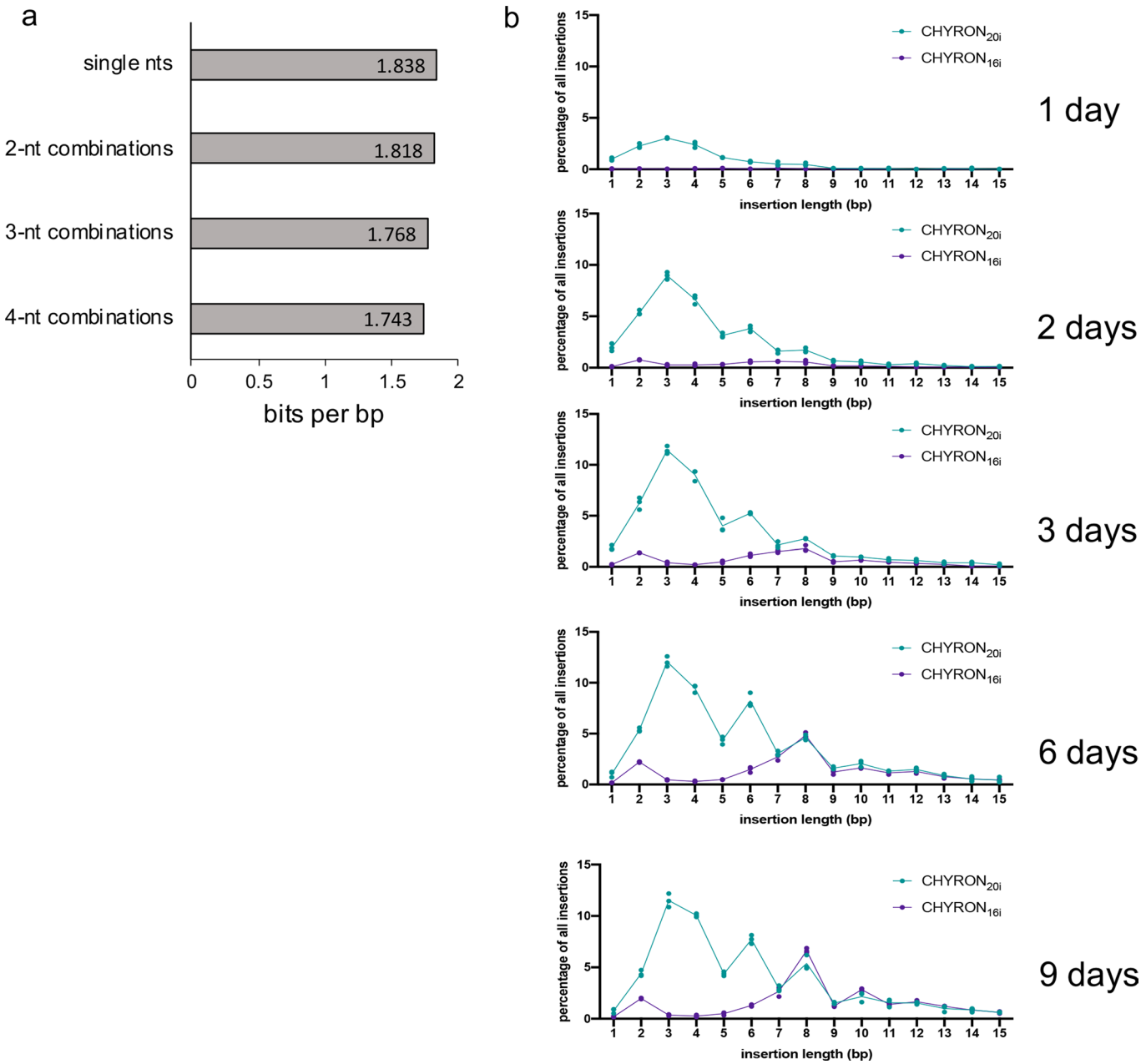
Extended Data Fig. 3 | TdT-mediated insertions are added 3 bp upstream of the PAM. For the pure insertions shown in Fig. 2a, the position of the insertion was determined, if the insertion sequence made this determination possible. Insertions were annotated as having an 'indeterminate' position, for example, if the 3′ nt of the insertion was identical to the protospacer nt 5′ of where the insertion was placed. Each point represents the mean of two technical replicates of a single biological replicate.



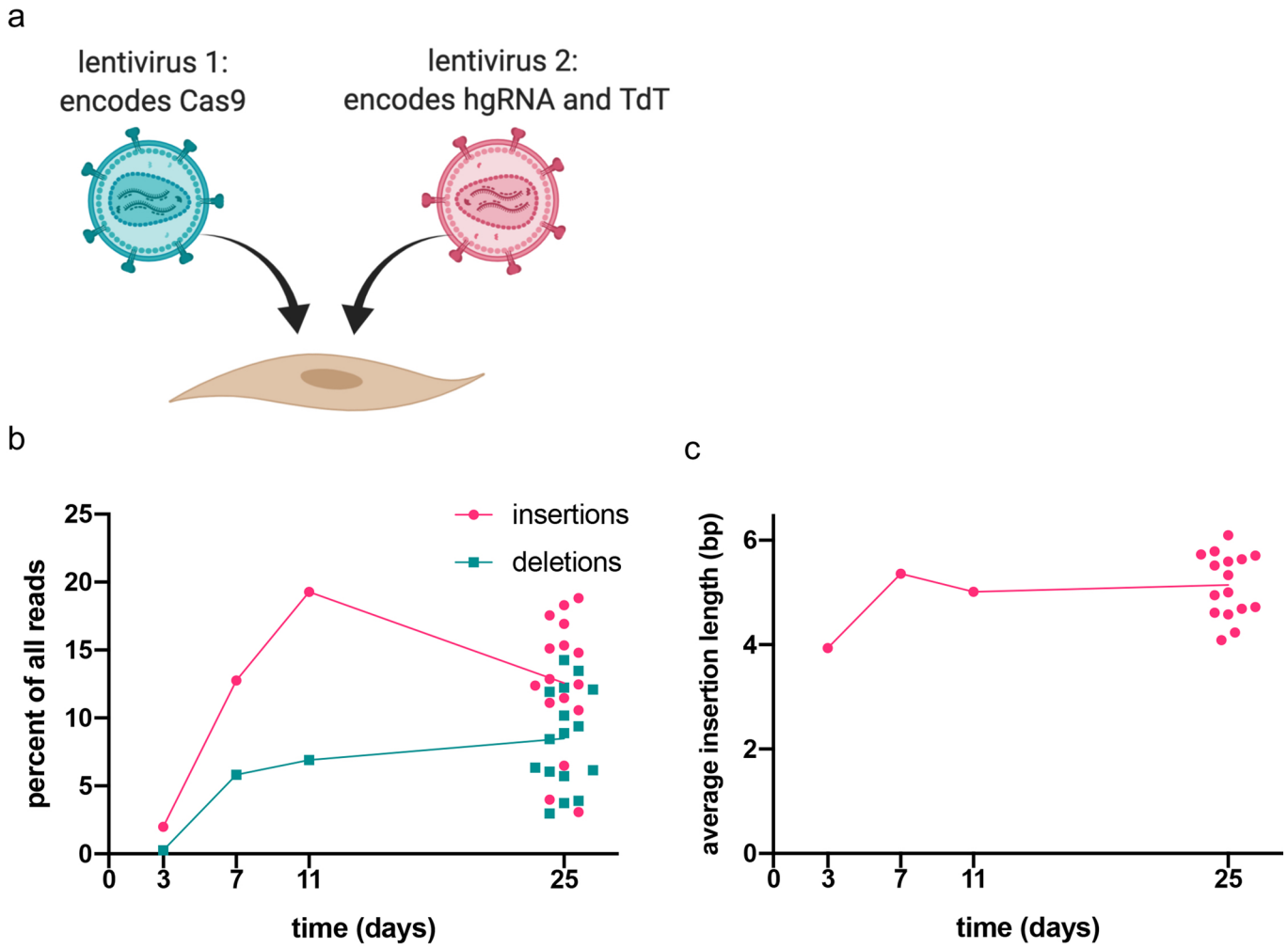
Extended Data Fig. 4 | Further characterization of CHYRON$_{20}$. a, Schematic of 293T-CHYRON$_{20}$. **b**, Western blots of samples shown in Fig. 3a–c. **c**, The Shannon entropy was calculated at each timepoint of this experiment.



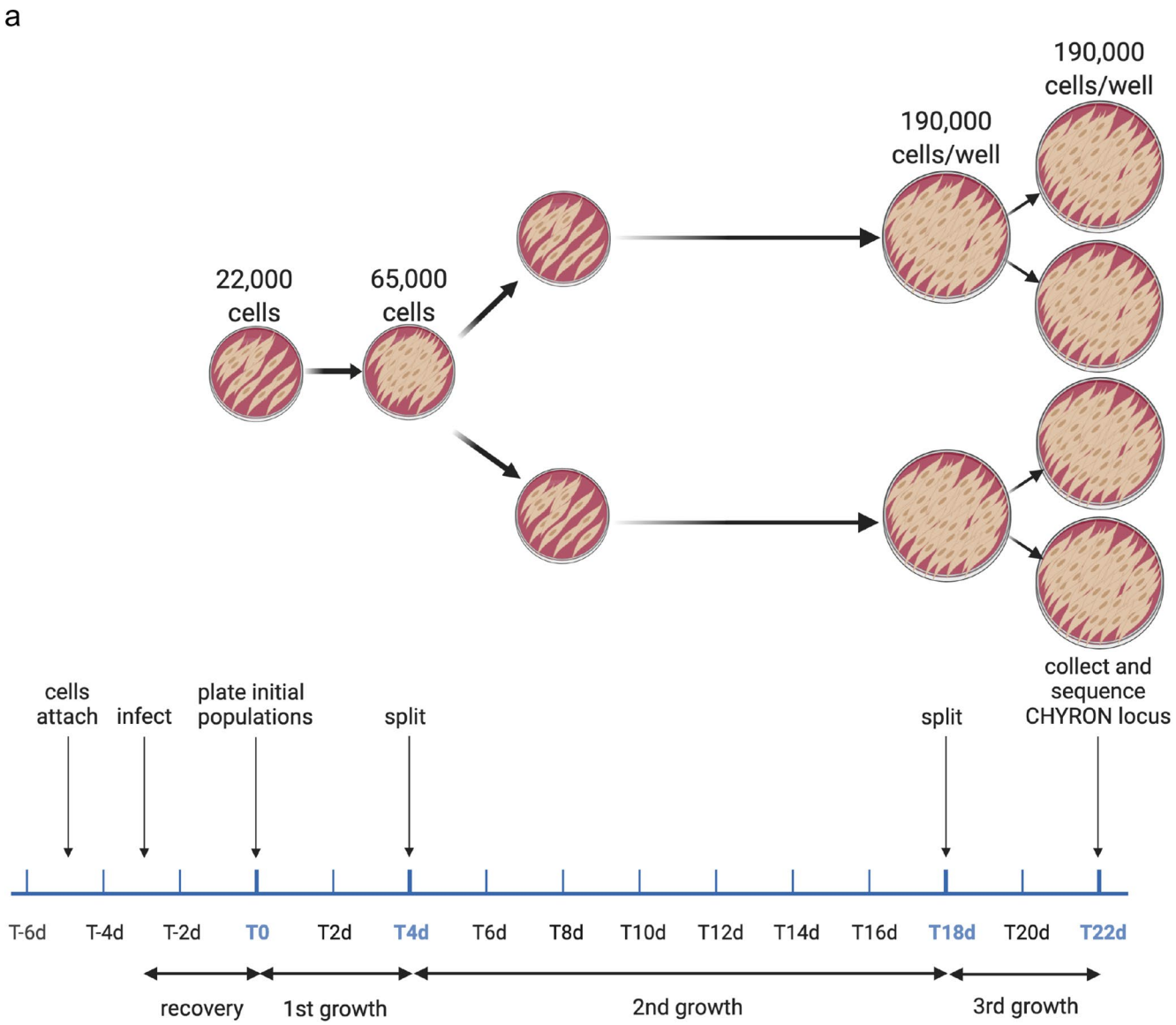
Extended Data Fig. 5 | Further characterization of CHYRON$_{20}$ over successive rounds of activity. a, Plan of the experiment for **b**, Fig. 3d, and Supplementary Fig. 3. **b**, Cas9 and TdT mediated multiple rounds of editing on an integrated hgRNA. The 293T-CHYRON$_{20}$ cell line was transfected with Cas9 and TdT to induce insertions, then single colonies isolated. The clonal isolate shown bears the insertions CTGAAAAACT and CCT. This cell line was treated and sequences analyzed as in Fig. 3d. Editing outcomes were determined to be the root CHYRON$_{20}$ sequence (not shown), deletions, both dominant insertions (not shown), any insertion containing these insertions or a shortened version as a prefix (CTGAAAAACT+, CCT+, CTGAA+), any insertion containing the sequences CTGAA or CCT other than as a prefix (*CTGAA* and *CCT*), or other insertions. Each point represents a single biological replicate.

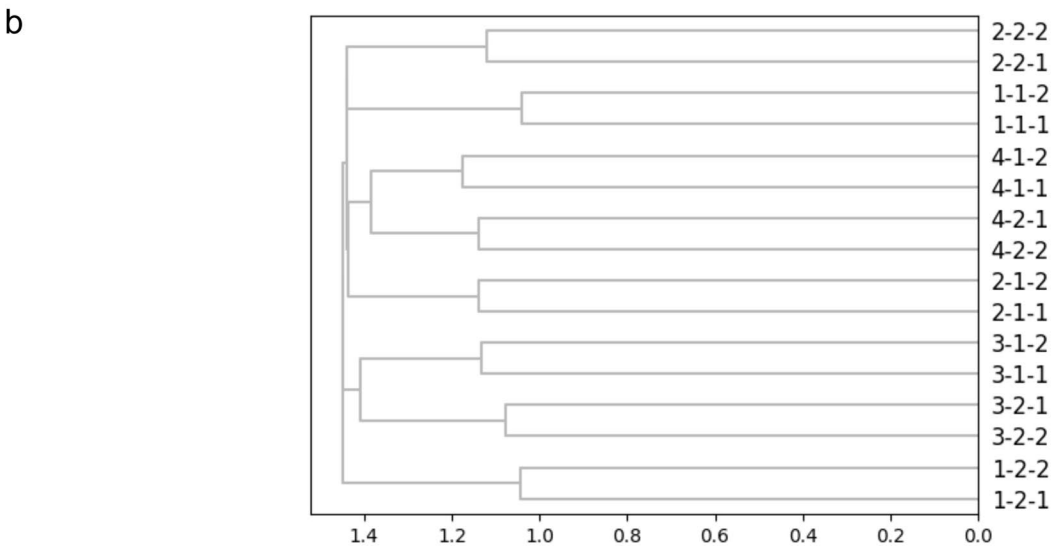


Extended Data Fig. 6 | Further characterization of CHYRON$_{20i}$ and CHYRON$_{16i}$. a, From the insertions detected in the experiment shown in Fig. 5, the proportions of all possible single-nt, 2-nt, 3-nt, and 4-nt sequences were determined, and the Shannon entropy calculated. The Shannon entropy was divided by the length of the sequences considered to calculate the bits per bp encoded in the insertions. **b**, CHYRON loci with an initial hgRNA length of 20 nt accumulated insertions quickly, with a gradual increase in length, whereas those with an initial hgRNA length of 16 nt accumulated insertions more slowly, ending up with a much longer insertion distribution. In the experiment shown in Fig. 4, 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 3 days. The CHYRON locus was analyzed by NGS and each sequence was annotated as root, pure insertion or any sequence that involves a loss of information. The percentage of total sequences that are a pure insertion of the indicated length was plotted. Each point represents a single technical replicate. (Some points overlap; three replicates were assayed.).



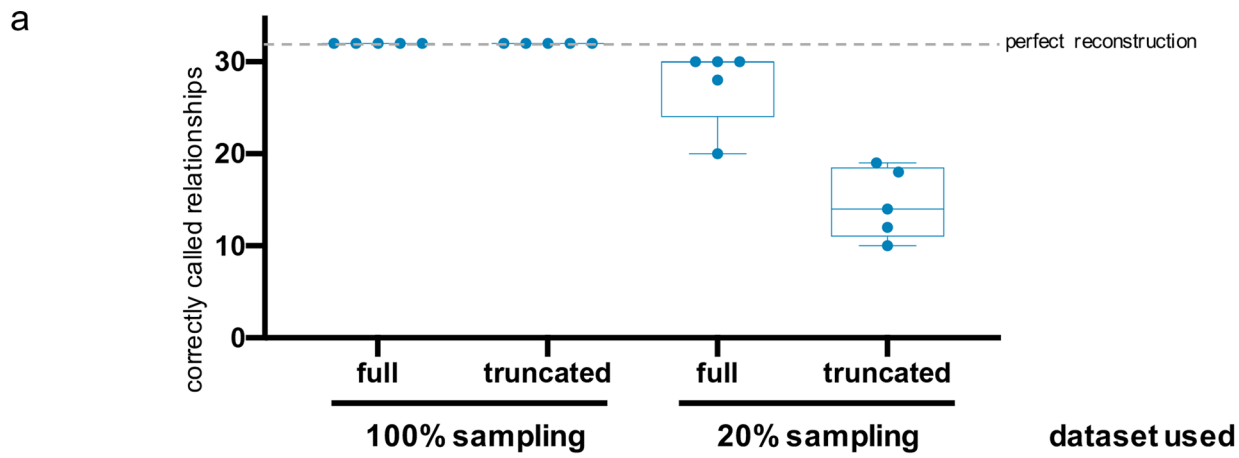
Extended Data Fig. 7 | CHYRON$_{17}$ delivered to primary cells by virus accumulates insertion mutations. a, Plan of the experiment. Cells were infected with two lentiviruses, one expressing Cas9, at a multiplicity of infection (MOI) of 15, and one expressing TdT and an hgRNA with a 17 nt initial spacer length, at an MOI of 30, then grown for the indicated time before collection. **b**, The CHYRON locus was analyzed by NGS and each sequence was annotated as in Fig. 2a. **c**, From these data, the average length of all pure insertions was calculated. For **b** and **c**, each point represents a single biological replicate.

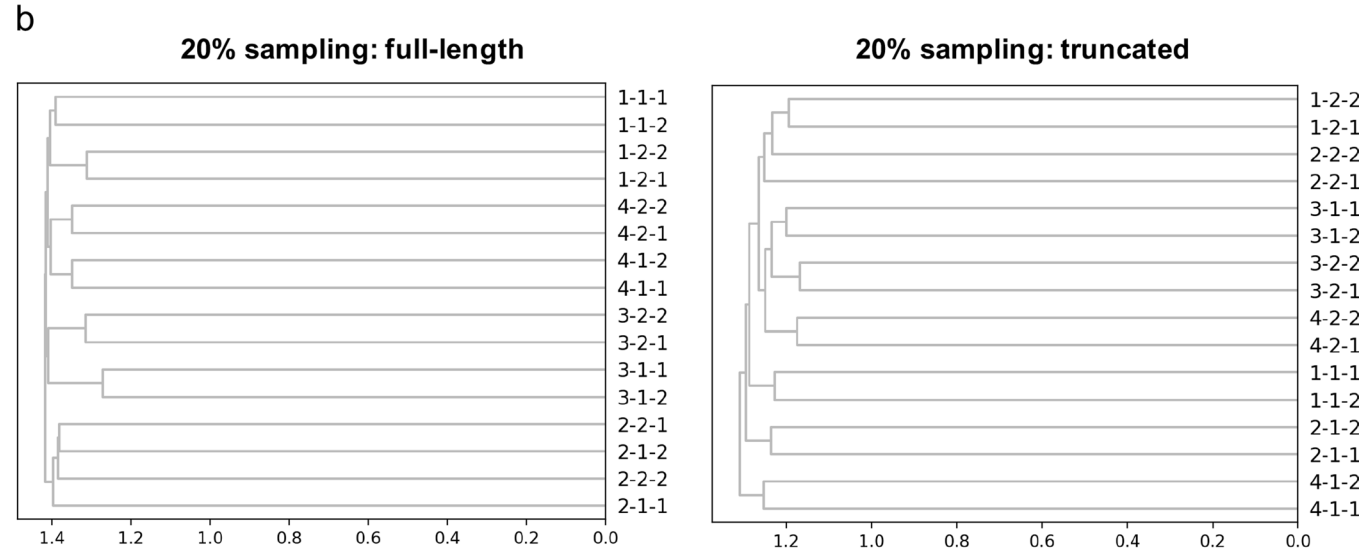


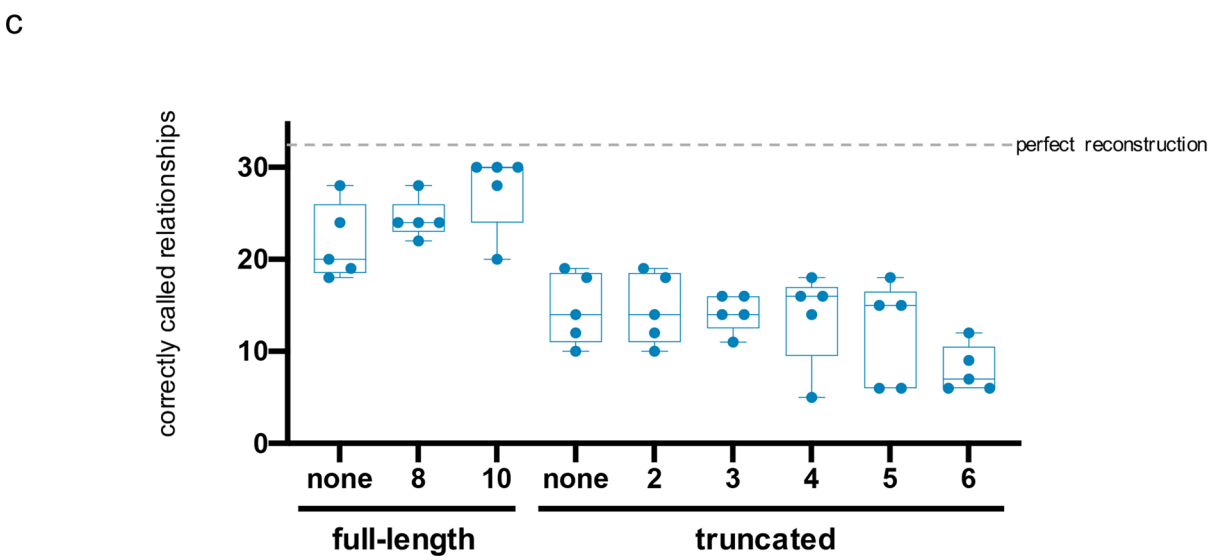


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Reconstruction of cell relatedness by DNA recording in primary cells. a, Plan of the experiment. This procedure was performed in quadruplicate to generate 16 final wells. 76,000 human adult primary dermal fibroblasts growing in 2 wells of a 24-well plate were infected with lentiviruses carrying $CHYRON_{17}$ at high MOI as described in Extended Data Fig. 7, then re-plated after 3 days to begin the experiment. Each well was split evenly into two new wells after 4 days, then split again after 14 more days, then collected 4 days after that. The CHYRON locus was sequenced, and unique insertions enumerated for each well. Low-abundance, artifactual insertions were removed (see Supplementary Fig. 4b and Methods). Lineage reconstruction was performed as in Fig. 5. **b**, Dendrograms for reconstruction using all unique insertions 7-15 bp in length.

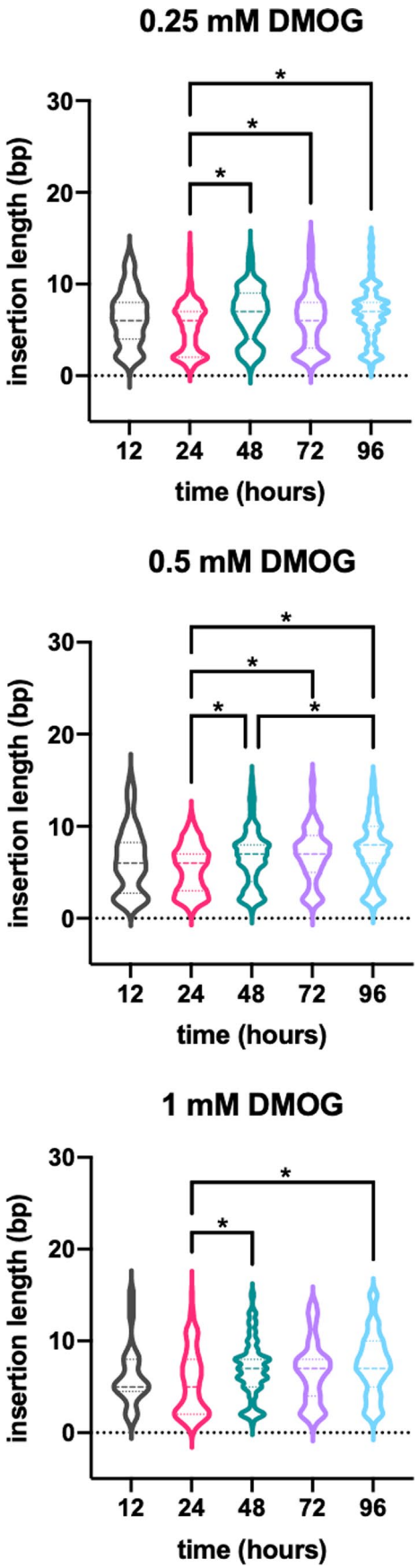






Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Reconstruction of cell relatedness by DNA recording requires high information when sampling is limited. For the experiment shown in Fig. 5, lineage reconstruction under sampling constraints would have been unsuccessful with a recorder that makes use of an hgRNA and Cas9 only. For each population in the experiment, the data set was degraded at random so that only 20% of all unique insertions remained. Then, each insertion was truncated so that the proportions of insertions encoding each amount of self-information matched the proportions of mutated hgRNA sequences encoding that amount of self-information in a published dataset. This pipeline was run five times and the number of correctly reconstructed relationships was calculated in the following way: for each well, the reconstruction was awarded one point for grouping the well with the proper sibling well and one point for grouping the well with at least 2 of the 3 other wells in its clade. Because the relationships among 16 wells were reconstructed, a maximum of 32 points is possible. **a**, Each point represents a replicate of the entire truncation and degradation process. For each reconstruction, the insertion length cutoff that produced the most accurate reconstruction for that sample was used. **b**, Representative dendrograms for reconstruction from 20% sampling data before or after truncation. For reference, the reconstruction on the left had a score of 30, and the reconstruction on the right had a score of 18. **c**, Reconstructions were scored for each minimum insertion length (in bp) used for reconstruction, which is noted below each bar. Results were plotted as in (**a**).



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | 293T-CHYRON$_{16}$ insertions grow longer with increasing duration of exposure to DMOG. For the experiment shown in Fig. 6d, the length of each insertion recorded in each condition was tabulated and plotted. At each dose, the timepoints were significantly different by one-way 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 ($p < 0.001$).

nature research

Corresponding author(s): Chang Liu

Last updated by author(s): Jan 30, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Only commercially-available software associated with Illumina HiSeq 2500 and MiSeq sequencers was used.
Data analysis	We used custom Python 2.7-3.7 scripts available at www.github.com/liusynevolab, which make use of the following Python packages associated with Python 3.7: sys, os, pickle, pprint, re, subprocess, glob, collections, motility, multiprocessing, numpy, matplotlib, regex, pandas, itertools, scipy, operator, tkinter, and csv; PEAR 0.9.10, available at https://cme.h-its.org/exelixis/web/software/pear/, Mapp, available at http://www.rle.mit.edu/sbg/resources/stgRNA/, Microsoft Excel for Mac version 16.45, and IBM SPSS version 25.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The list of figures associated with NGS data sets is in Table S1. All NGS data sets have been deposited in the Sequence Read Archive, accession #PRJNA561027.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were chosen to detect and limit distortions introduced in the NGS library prep process (e.g., PCR jackpotting). Because technical replicates had similar characteristics, these types of distortions were not present.
Data exclusions	Lineage reconstruction data from an experiment in which sampling was quite poor were not included in the paper. This was a separate experiment, not shown at all, that was similar to that shown in Figure 5. However, the raw data files are listed in Table S1 and available in the Sequence Read Archive.
Replication	The results in Figure 2a-b and Extended Data Figure 3 reflect three biological replicates. Figure 2c reflects, in most cases, two biological replicates. The lineage reconstruction in Figure 5 is robust when different groups of 25% of the data were sampled at random (see Supplementary Figure 4e).
Randomization	In all cases, experiments started with an identical pool of cells, which were subsequently treated in different ways, so randomization was not required.
Blinding	For the lineage reconstruction described in Figure 5, the researchers (MWS and BSA) setting the initial reconstruction parameters did not know which wells were which. Similarly, during the library prep and initial lineage reconstruction for Extended Data Figure 8, the researcher (TBL) setting the reconstruction parameters did not know which wells were which. For all other experiments, blinding was not required, because analyses were straightforward and standard.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	We used Guide-it Cas9 Polyclonal Antibody (Clontech #632607), anti-TdT Antibody (Abcam #ab14772), anti-Actin Antibody (Abcam #ab14128), anti-Rabbit IgG (Sigma-Aldrich #A0545), and anti-Mouse IgG (R&D Systems #HAF007) .
Validation	The Cas9 antibody detects bands at the predicted sizes when various Cas9 fusion proteins are expressed (Figure S3d). The TdT antibody detects a band at the predicted size only when we express ectopic TdT (Figure S3a-b). The actin antibody detects a band at the expected size, and is widely used: see 29 references at https://www.abcam.com/actin-antibody-c4-ab14128-references.html#top-986.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	293T cells (catalog number CRL-3216) and human adult primary dermal fibroblasts (catalog number PCS-201-012) were obtained from ATCC. For the experiments shown in Extended Data Figure 2, we used PCS-201-012, lot number 70015617. For the experiments shown in Extended Data Figures 7-8, we used PCS-201-012, lot number 61683453.

Authentication	Cells were obtained directly from ATCC. In all cases, a PCR directed at the human genome produced a product of the expected sequence (see Figure 2 for 293T cells and Extended Data Figure 2 for primary dermal fibroblasts), suggesting the cells are human.
Mycoplasma contamination	293T 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, Inc.). Primary dermal fibroblasts were certified mycoplasma-free by ATCC and were used immediately upon receipt.
Commonly misidentified lines (See ICLAC register)	None were used.