Molecular recording of sequential cellular events into DNA

Theresa B. Loveless[1,2,3,#,*], Courtney K. Carlson[1,2,#], Vincent J. Hu[1,2,4], Catalina A. Dentzel Helmy[1,2], Guohao Liang[1,2], Michelle Ficht[1,2], Arushi Singhai[1,2], and Chang C. Liu[1,2,3,5,6,*]

[1]Department of Biomedical Engineering
[2]Center for Synthetic Biology
[3]NSF-Simons Center for Multiscale Cell Fate
[4]Graduate Program in Mathematical, Computational and Systems Biology
[5]Department of Chemistry
[6]Department of Molecular Biology and Biochemistry
University of California, Irvine, 92697, USA

[#]These authors contributed equally to this work.
*e-mail: ccl@uci.edu; theresa.berens.loveless@gmail.com

Summary: Genetically encoded DNA recorders noninvasively convert transient biological events into durable mutations in a cell's genome, allowing for the later reconstruction of cellular experiences using high-throughput DNA sequencing[1]. Existing DNA recorders have achieved high-information recording[2–14], durable recording[3,5–10,13,15–18], prolonged recording over multiple timescales[3,5,8,10], multiplexed recording of several user-selected signals[5–8,18], and temporally resolved signal recording[5–8,18], but not all at the same time. We present a DNA recorder called peCHYRON (prime editing[19] Cell HistorY Recording by Ordered iNsertion) that does. In peCHYRON, prime editor guide RNAs[19] (pegRNAs) insert a variable triplet DNA sequence alongside a constant propagation sequence that deactivates the previous and activates the next step of insertion. This process results in the sequential accumulation of regularly spaced insertion mutations at a synthetic locus. Accumulated insertions are permanent throughout editing because peCHYRON uses a prime editor that avoids cutting both DNA strands, which risks deletions. Editing continues indefinitely because each insertion adds the complete sequence needed to initiate the next step. Constitutively expressed pegRNAs generate insertion patterns that support straightforward reconstruction of cell lineage relationships. Pulsed expression of different pegRNAs enables the reconstruction of pulse sequences, which may be coupled to biological stimuli for temporally-resolved multiplexed event recording.

Architecture of peCHYRON

An ideal DNA recorder should act much like a news ticker, often displayed as a scrolling chyron at the bottom of television programs. A news ticker records and displays a continuously updating stream of events (**Figure 1a**). Critical properties include 1) information complexity: the amount of information that can be used to describe each event, 2) order, the ability to record events in the

sequence they arrive, 3) punctuation, the ability to delimit one event from the next in the record, 4) durability, the ability to record new events without corrupting or deleting the record of older events, and 5) continuous operation, the ability to record new events over long timescales. Existing DNA recorders, even ones based on progressive insertional mutagenesis, contain fundamental mechanistic tradeoffs that prevent them from achieving all these properties simultaneously[2–18], whereas the peCHYRON architecture we present here does not (**Figure 1b**).

In peCHYRON, a unique 20-bp locus in a mammalian cell's genome is targeted by a pegRNA. The pegRNA directs prime editor 2 (PE2)[19] to reverse transcribe the pegRNA's programmable template sequence into the locus. The reverse transcriptase (RT) template sequence is programmed to insert a variable 3-nt sequence that encodes up to 6 bits of information followed by a constant 17-nt propagator sequence that serves as the target sequence for the next insertion step. After one cycle of insertion, the previous propagator sequence is no longer PAM-adjacent, and therefore inactive, while the new propagator sequence that forms the new 20-bp target site becomes PAM-adjacent, and therefore active. The next step of editing inserts another variable 3-nt sequence along with the next propagator sequence, iterating indefinitely. In this manner, each variable 3-bp sequence, which we refer to as a signature mutation, records an event. The information content of the signature mutation can be up to 6 bits, representing sufficient complexity to theoretically record 64 different types of events. New signature mutations appear in the order of events recorded. Constant 17-bp sequences separate signature mutations, acting as punctuation. The recording process does not change previous records as edits occur through sequential insertions, and recording can propagate continuously without end (**Figure 1b**).

The simplest implementation of peCHYRON would involve the sequential insertion of a constant propagator sequence (**Extended Data Figure 1a-b**). However, iterative addition of a single propagator sequence requires a pegRNA whose RT template sequence (containing the 17-nt

propagator sequence to be inserted) and primer binding sequence (PBS)[19] share identity. During the RT step of prime editing, this would allow primer binding in the pegRNA's RT template sequence rather than the PBS, resulting in no net insertion. We therefore settled on the use of two alternating 17-bp propagator sequences in our peCHYRON design so that the RT template sequence and PBS in each pegRNA are distinct. The peCHYRON architecture we implemented is shown in **Figure 1c** where one pegRNA targets site A and adds B (A->B) and the other pegRNA targets site B and adds A (B->A), allowing for propagation. Recorded information in the resulting peCHYRON locus is easily interpretable (**Figure 1d**)

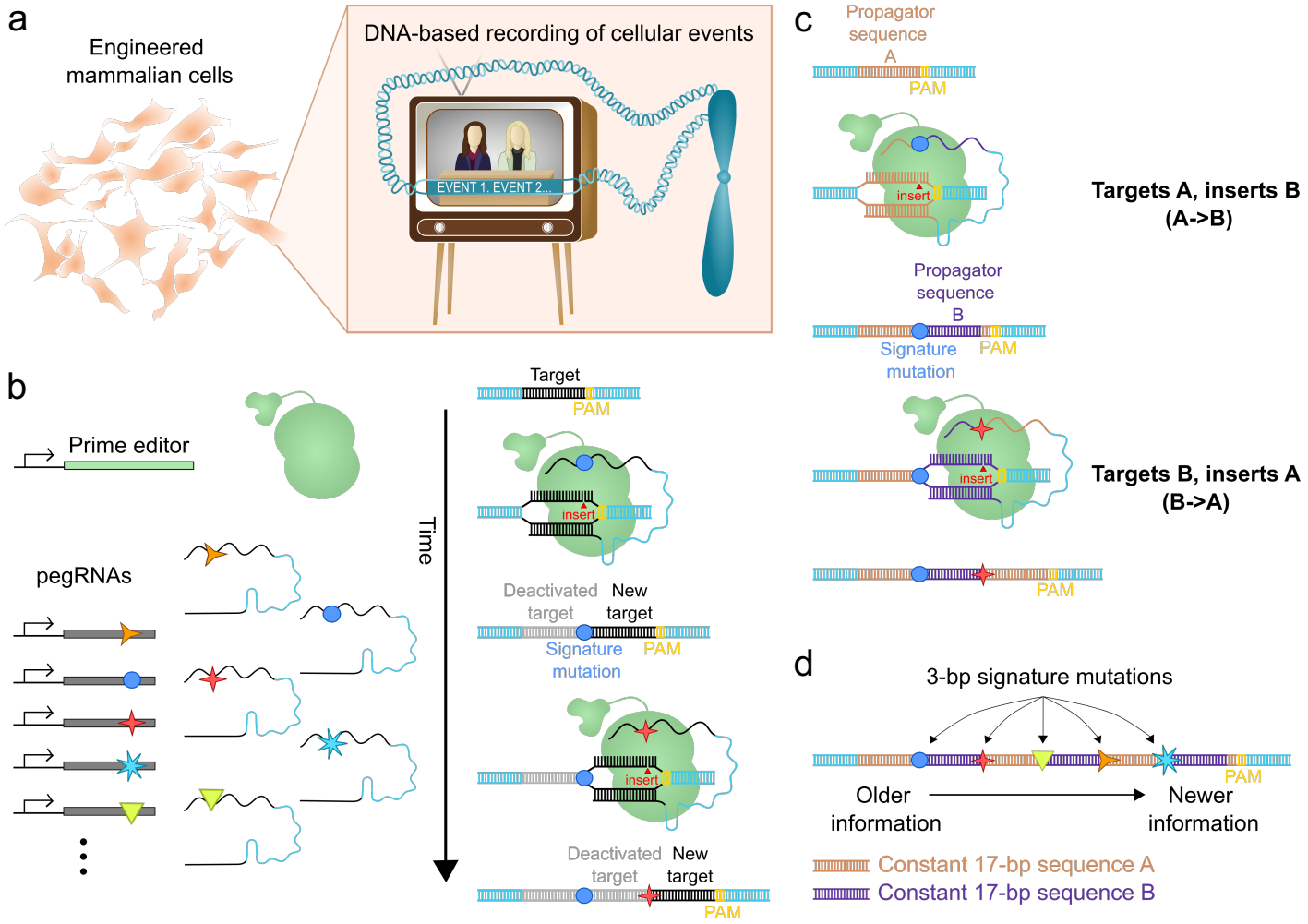


Figure 1. Overview of peCHYRON. a, Using peCHYRON, transient events are recorded into DNA in the order they occur, as in a scrolling chyron. **b**, Iterative sequential insertion of signature mutations via prime editing. Signature mutations are variable 3-bp sequences, each theoretically encoding up to 6 bits of information. A collection of pegRNAs facilitates the insertion of signature mutations at the recording locus during continuous and iterative editing. In a single cycle of prime editing, a signature mutation and a 17-bp propagator sequence are inserted at the recording locus adjacent to the PAM. The newly-inserted propagator sequence forms the target site for the next cycle of editing because it is now PAM-adjacent. The previous target site is deactivated due to separation from the PAM. This process can repeat indefinitely, resulting in ordered accumulation of regularly-spaced signature mutations. **c**, Propagation with alternating pegRNAs. To avoid issues with repetitive sequence elements, the recording locus alternates between addition of two different 17-bp propagator sequences. In one step, a pegRNA targets sequence A and inserts sequence B. In the next step, a pegRNA targets sequence B and inserts sequence A. The process repeats and continuous propagation of the recording locus proceeds without end. **d**, How to interpret information at a peCHYRON locus. Information-rich signature mutations are separated by constant 17-bp propagator sequences, resulting in clear punctuation between sequential editing events. The age of a signature is reflected by its position in the record, with PAM-proximal signatures reporting on the most recent events.

Identification of efficient parts for peCHYRON

In the original report of prime editing by Anzalone *et al.*[19], it was found that an 18-bp sequence ("6xHis") could be inserted at target genomic sites using PE3, a prime editor that achieves high genome-editing activity by nicking both DNA strands. Since the core mechanism of peCHYRON involves the insertion of a 20-bp sequence (*i.e.* a 3-bp signature mutation and a 17-bp propagator sequence), PE3's ability to insert 6xHis was a natural starting point in the design of peCHYRON. Several developments were necessary.

First, we reasoned that the modest rates of unintended insertion or deletion mutations (indels) associated with PE3[19] could be problematic in peCHYRON, because spurious indels at any step of the iterative insertion process would terminate recording. We therefore tested whether 6xHis could be inserted using PE2, which does not cut both strands of target DNA[19] (**Figure 2a**). To our surprise, we found that 6xHis was inserted by PE2 at comparably high efficiency as by PE3. As expected, the 6xHis insertion made by PE2 was accompanied by a much-lower rate of unintended insertion or deletion mutations (indels) (**Figure 2a**). Consequently, we decided to use PE2 as the prime editor for peCHYRON.

Second, we tested whether the 18-bp 6xHis insertion sequence identified in Anzalone *et al.*[19] could be converted into a 20-bp insertion sequence in which the first 3 bps constituting the signature mutation are randomizable (**Figure 2b**). A collection of pegRNAs, each of which targets a defined site in the genome ("site 3"[19,20]) and contains the template for inserting a distinct 3-nt signature mutation followed by a constant 17-nt sequence nearly identical to 6xHis, was tested for PE2-mediated editing in HEK293T cells. Although some variation was observed, most sequences were inserted with high efficiency similar to that of the original 18-nt 6xHis sequence.

Third, we attempted to make a pegRNA that uses the inserted 6xHis-based sequence as its target, necessary for subsequent propagation. Recall that the 6xHis-based sequence was inserted at the site3 target. If we consider site3 to be A and 6xHis to be B (see **Figure 1c**), the original pegRNA that installed 6xHis corresponds to the A->B step and a pegRNA designed to insert site3 at the 6xHis target corresponds to the B->A step. We designed a B->A pegRNA that inserts site3. However, we did not observe efficient editing (**Extended Data Figure 1c**). We also did not observe efficient insertion of any of 20 arbitrary sequences at the 6xHis target, suggesting that successful insertion sequences are rare (**Extended Data Figure 1d**). We therefore performed a high-throughput screen in which a library of ~100,000 pegRNAs, each designed to target the 6xHis site and add a random 20-nt sequence, was transfected into 293T cells that contained the target 6xHis site in their genomes ($293T_{6xHis}$). After 3 days, we sequenced the target sites to identify enriched insertions, which we then validated individually for high insertion activity. Efficient insertion sequences were uncommon, but several were identified (**Figure 2c** and **Extended Data Figure 1e**). For the ten most efficient insertion sequences, we subsequently optimized PBS length and ensured they could tolerate variable signature mutations (**Extended Data Figure 2a**).

For propagation to occur, an inserted sequence must then act as the target for the next step. Of the ten efficient insertion sequences, we identified two that tolerated variable signature mutations and behaved as good targets for the insertion of 6xHis sequences containing signature mutations (**Extended Data Figure 2a-b**). Let us call these two sequences B_4 and B_7 and the 6xHis sequence A (see **Figure 1c**). The downselection from the screening steps described above ensured that B_4 and B_7 can successfully act with A both in the A->B step (*i.e.*, target A and insert B_4 or B_7) and the B->A step (*i.e.*, target B_4 or B_7 and insert A), providing the necessary ingredients for peCHYRON. When we simultaneously transfected A->B and B->A pegRNA pairs for both B_4 and B_7, we observed extended propagation (**Extended Data Figure 2c**). One pair resulted in slightly

more efficient propagation (**Extended Data Figure 2c**), so we performed all subsequent experiments with that pair.

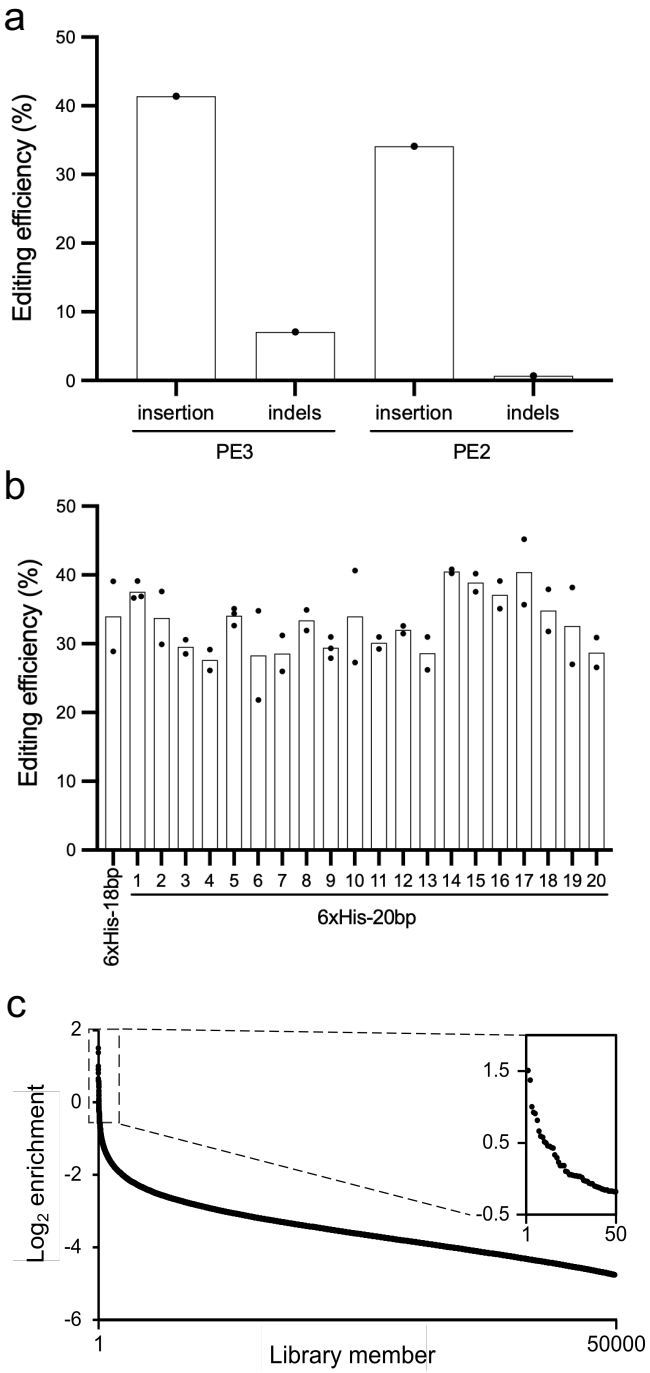


Figure 2. Identifying efficient components for iterative editing with two alternating propagator sequences. a, PE2, a prime editing system that does not make double-strand breaks, efficiently adds insertions at the peCHYRON locus. An 18-bp 6xHis sequence was inserted at site3 in HEK293T cells using PE3 and PE2, and the rates of intended insertions vs. spurious indels were assessed. **b,** 20-bp sequences derived from the 18-bp 6xHis sequence were inserted efficiently. Each 20-bp sequence was nearly identical to the original 6xHis sequence, but with the first 3 bp randomized (3-bp signature + 17-bp propagator sequence). Insertions were made at site3 in HEK293T cells using PE2. Points represent biological replicates. Efficiencies from different experiments were normalized using samples that were present in each experiment (sequences 1, 5, and 9). **c,** Efficient pegRNAs that target 6xHis and insert a different 20-bp sequence (A->B) were rare. A high-throughput screen was used to identify pegRNAs that could alternate with the 20-bp variant of the 6xHis tag to iteratively edit the peCHYRON locus. In the screen, a library of ~100,000 pegRNAs with variable insertion sequences was transfected into HEK293T cells along with PE2. High enrichment values correspond to pegRNAs that frequently edited the 6xHis target site without being over-represented in the original transfection mix. The top 50,000 enriched library members were plotted, and the inset shows the top 50.

Sequential editing results in regularly-spaced ordered insertions

Having identified a proper A->B and B->A pair of pegRNAs that could propagate continuously, we created a full peCHYRON system and tested its performance characteristics. To do so, we made a set of 14 PiggyBac Transposase cargo vectors that contain PE2 along with subsets of 21 pairs of A->B and B->A pegRNAs (**Extended Data Figure 3**). Each A->B pegRNA contained one of 21 different signature mutations in the B sequence and each B->A pegRNA contained one of 20 different signature mutations in the A sequence. We created and passaged a cell line, 293T-peCHYRON$_{PB14v1}$, in which these cargos were stably integrated. We observed that the recording locus accumulated insertions over time following the expected pattern (**Figure 1d**), generating ordered strings of insertions that recorded up to 7 signatures over the moderate 13-day timecourse tested (**Figure 3a-b**). The accumulated insertions represented ample information: considering the relative frequencies of signature mutations observed, we calculated a Shannon entropy[21] of 3.2 bits per signature mutation (**Supplementary Table 1**), which compares favorably to the average information content of a letter in English[22]. Importantly, 99% of sequences observed had no spurious indels (**Figure 3c**), suggesting that each propagation step accurately inserted the intended sequence without corrupting previous insertions. The rate of editing observed for early insertion events was maintained in later insertion events (**Figure 3d**), suggesting that propagation efficiency does not diminish over sequential editing. peCHYRON therefore avoids the common failure modes of existing DNA recorders, unscheduled mutations[2,4,11,14] and declining efficiency[3,9,10,13–18]. Preventing such failures opens up new possibilities in DNA recording.

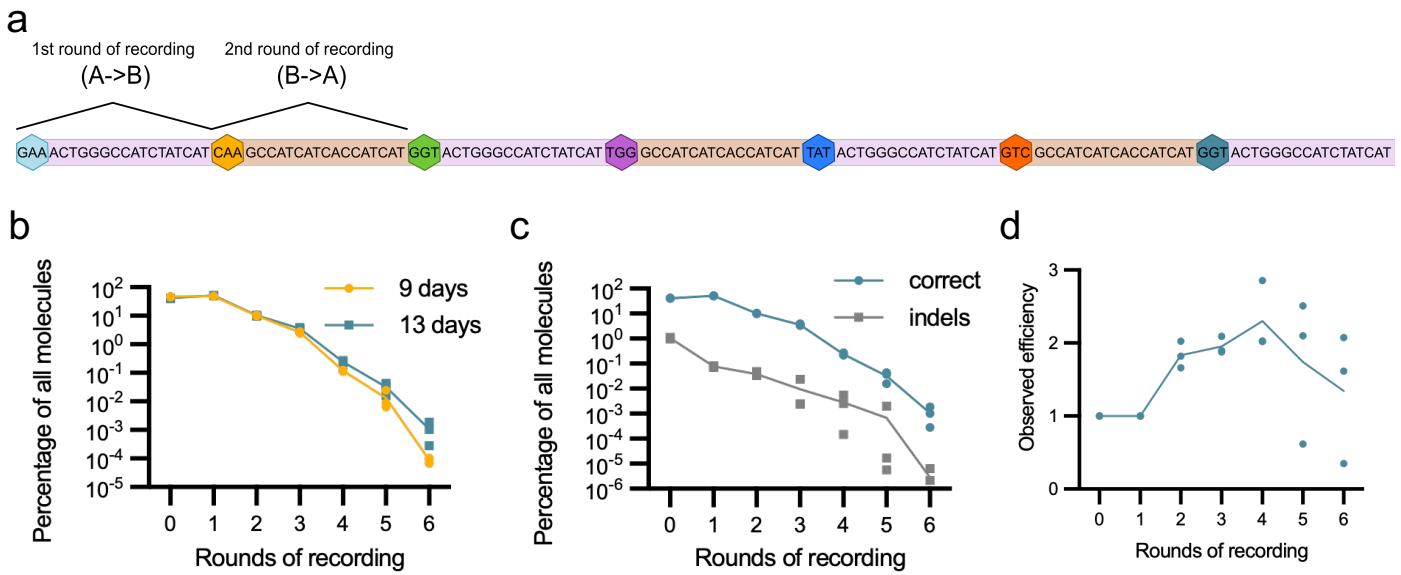


Figure 3. Characterization of peCHYRON performance. a, peCHYRON recording propagated for many rounds. Example high-throughput sequencing read with 7 rounds of sequential insertion is shown. Constant 17-bp propagator sequences B and A are highlighted in lilac and tan, respectively, while 3-bp signature mutations are highlighted in colorful hexagons. The first two rounds of sequential insertion are labeled in 20-bp sections. **b,** Insertions accumulated during propagation with the most efficient pair of A->B and B->A pegRNAs. Expression cassettes for PE2 and 21 versions of each pegRNA (21 signatures for A->B and 20 for B->A) were co-transfected with PiggyBac transposase for stable integration into the genome of $293T_{6xHis}$ and cells were grown for the indicated time after the initial transfection. Insertions of 20 bp were considered to correspond to 1 round of insertion, 40 bp to 2 rounds, and so forth. Each point represents one technical replicate. Three replicate values are shown for each round; some points overlap. **c,** The rate of spurious indel mutations throughout the passaging experiment was very low. For the 13-day timepoint, each read was assigned to a bin according to its size; insertions that exactly match the expected sizes were considered "correct," and all other mutations were considered "indels." Each point represents one technical replicate. Three replicate values are shown for each round; some points overlap. **d,** Editing rates did not diminish throughout rounds of sequential insertion. The efficiency for each cycle of peCHYRON recording was compared to the efficiency that would be expected if the editing rate observed for the first two rounds of recording stayed constant over time. Observed efficiency was normalized to the expected efficiency, so a value of 1 indicated that the observed efficiency matched the expected efficiency. Each point represents one technical replicate. 3 replicate values are shown for each round; some points overlap.

Reconstruction of lineage relationships using peCHYRON

We applied peCHYRON to trace the lineage relationships among populations of cells descended from a parent population *via* a complex splitting process (**Figure 4a**). We transiently transfected one population of $293T_{6xHis}$ cells with plasmids encoding PE2 and 42 pegRNAs representing A->B and B->A pairs with varied signature mutations. After allowing the cells to grow for 8 days, we isolated 4 populations of 20,000 cells each. The next day we transfected again, allowed the cells to grow for 3 days, then split each population to yield 8 populations. Cells were transfected again the next day, then allowed to grow for 4 additional days, before being split again to yield 16 final populations. Two days later, ~1.6 million cells were collected from each population, DNA was extracted and the peCHYRON recording locus was subjected to amplicon sequencing. Three of the 16 populations were excluded from analysis so that the researchers performing the reconstruction could be blinded to both the identity of the wells and the shape of the lineage tree. From the sequences in the remaining 13 populations, we first filtered sequences by length, then by frequency of occurrence. We discarded all sequences with fewer than 4 signature mutations as the probability of generating a specific 3 signature mutation sequence by chance rather than by descent was significant. Finally, we discarded sequences with frequencies of occurrence below a cutoff determined by the kneedle algorithm[23]. This represented a filter that removed spurious sequences that are likely library prep artifacts. After these two filtering steps, a total of 4571 sequences remained (**Supplementary Table 2**). For each pair of wells, we counted the number of shared identical peCHYRON sequences to calculate Jaccard similarity[24] and then used the Jaccard similarity scores to perform agglomerative hierarchical clustering. The resulting tree accurately reconstructed all aspects of the splitting procedure (**Figure 4b**).

Calculating Jaccard similarity worked well for comparing populations of cells, but it only takes into account identical peCHYRON sequences shared among the populations. Because the peCHYRON recording locus progressively acquires signature mutations in temporal order, related cells can share initial edits and then diverge. Sequences with shared early signature mutations but diverged late signature mutations contain higher resolution information about lineage relationships, theoretically approaching single-cell-resolution. We sought to establish a lineage reconstruction algorithm that would take advantage of this ordered nature of recording. To do so, we modified the Jaccard similarity index to allow for partial matches. Pairs of populations are compared as follows. First, each sequence of signature mutations is split into a collection of all possible subsequences that start from the first signature mutation (*i.e.* prefixes). Each prefix generated from a sequence is given a fractional weight equal to the inverse of the number of prefixes making up the sequence. This results in a multiset of prefixes. We then computed a weighted multiset Jaccard similarity between all pairs of multisets. To illustrate this, consider a record ABCD where A, B, C, and D each represent a unique 3-bp signature mutation. This is split into 4 prefixes, A, AB, ABC, and ABCD, each given a weight of ¼. Now, consider a second record, ABCEF. This is split into 5 prefixes, A, AB, ABC, ABCE, and ABCEF, each given a weight of ⅕. When these are compared, the prefixes A, AB, and ABC match, and we add the minimum count, ⅕, for each match. The sum of these matches becomes the numerator for the Jaccard similarity, and the denominator is simply the sum of the cardinalities of the sets minus the numerator. With this new algorithm, the full splitting procedure was again accurately reconstructed (**Figure 4c**). Lineage reconstruction outperformed that done with Jaccard similarity at almost every level of random downsampling and gave near-perfect reconstructions even when downsampling to only 20% of the data (**Extended Data Figure 4**), forecasting the utility of this reconstruction algorithm in realistic lineage tracing experiments where populations are poorly sampled.

Reconstruction of pegRNA pulse sequences using peCHYRON

Each signature mutation in the peCHYRON recording locus reflects the identity of an exact pegRNA. Since recording is sequential, the order of signature mutations in the resulting locus is the order of pegRNAs used. This makes it possible to decode pegRNA pulse sequences. Additionally, different pegRNAs can be expressed from inducible promoters of interest, enabling multiplexable reconstruction of the temporal sequence of induction events corresponding to specific biological signals of interest. To test whether the order of signature mutations could be used to reconstruct temporally-resolved histories of transient events, we used populations of $293T_{6xHis}$ cells to record pegRNA pulse patterns (**Figure 4d**). For each population, we used a random number generator to choose a pair of A->B and B->A pegRNAs marked by a unique pair of signature mutations. We then transiently transfected plasmids encoding the chosen pegRNA pair along with PE2 and grew the cells for 13 days. We then randomly chose a different pair of A->B and B->A pegRNAs marked by a unique pair of signature mutations, transfected cells again, and allowed the cells to grow for 9 days. We collected cells at the end of the experiment, extracted DNA, and subjected the peCHYRON locus to amplicon sequencing. The resulting sequences accurately reflected the patterns of transfection. **Figure 4e** shows the peCHYRON loci that, for simplicity, contain exactly 4 signature mutations and the proportion of sequences at each of the 4 positions bearing each signature. The temporal sequences of pegRNA transfection pulses are immediately apparent from the order of signature mutations observed. This suggests that peCHYRON will be able to accurately reconstruct the timing of the many biological phenomena that take place over ~1 week (*e.g.*, the epithelial-mesenchymal transition[25]) when pegRNA pulses are linked to biological stimuli.

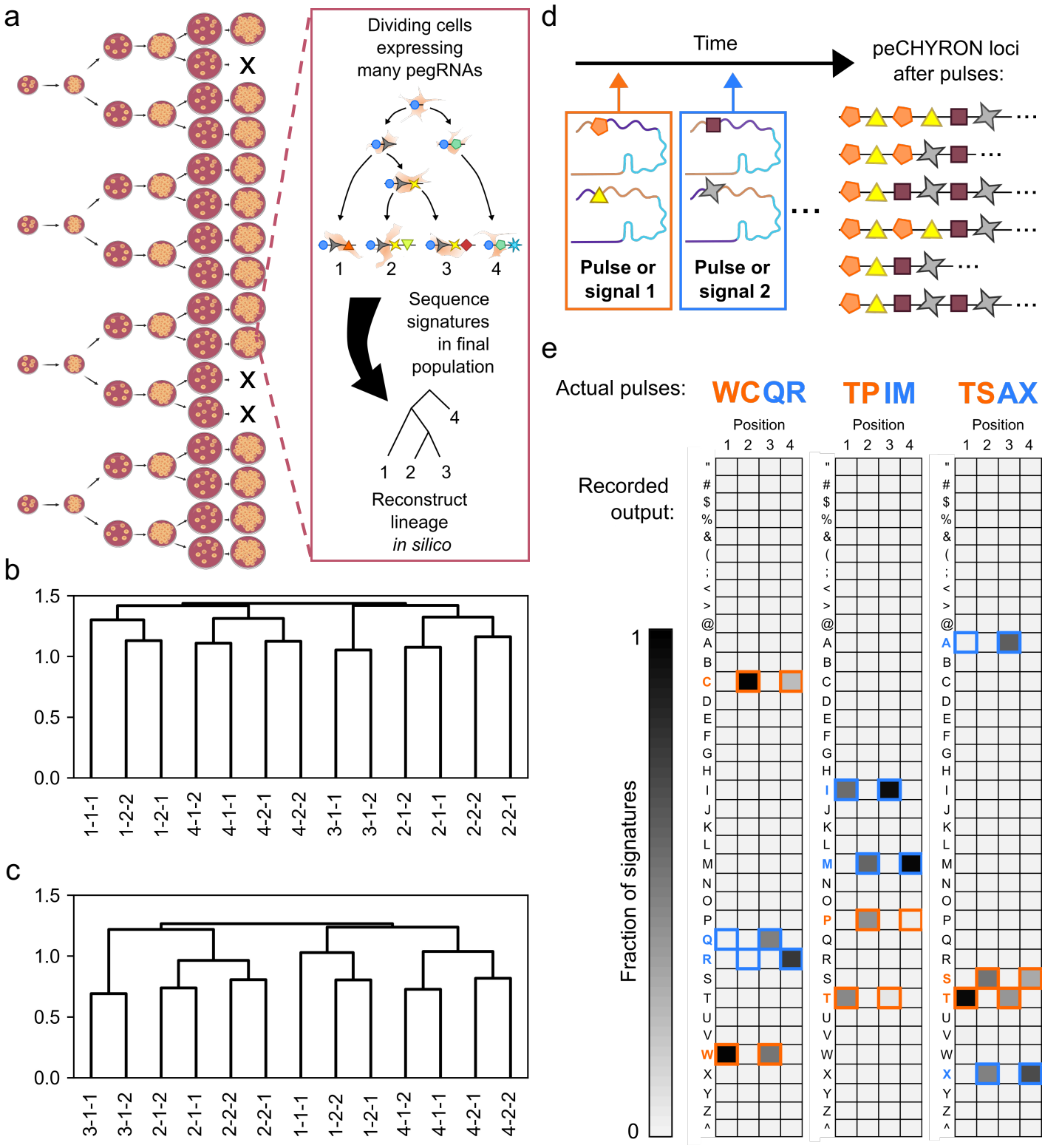


Figure 4. peCHYRON enables lineage tracing and temporally-resolved event recording.

Figure 4. peCHYRON enables lineage tracing and temporally-resolved event recording. a, Culture splitting procedure for lineage tracing experiment. Inset shows how signature mutations accumulate in individual cells, then inform lineage reconstruction. **b,** Culture splitting patterns were accurately reconstructed using the Jaccard similarity index. **c,** Culture splitting patterns were accurately reconstructed using a modified Jaccard similarity index that accounts for sequences that share early signature mutations but diverge in later signature mutations. **d,** General approach to recording pulses (or any time-dependent signals) that are linked to expression of pegRNAs with known signature mutations. A pair of A->B and B->A pegRNAs are expressed together for any signal to be recorded, resulting in alternation between two signature mutations. Colored shapes represent signature mutations. **e**, Heatmaps showing enrichment data for samples transfected with PE2 and different pairs of A->B and B->A pegRNAs at different times. Signature enrichment was calculated at each position in peCHYRON loci harboring precisely 4 edits. 1-letter symbols along the y-axes represent 3-bp signature mutations. The analysis pipeline searched for all signatures that the random number generator used to choose which pegRNAs to transfect could have chosen. Signatures colored in orange and blue signify the first and second pairs of sequentially transfected pegRNAs, respectively. Orange and blue boxes outline the possible positions for each signature. Because alternating propagator sequences were used, individual pegRNAs can only edit odd or even positions, not both.

Discussion

Our results establish peCHYRON as an advanced DNA recorder that autonomously generates ordered, high-information records *in vivo*. These records are exceptionally durable, can be parsed to decode the temporal order of recorded events, and can be arbitrarily long since the rate of sequential insertions in peCHYRON remains constant throughout recording. In keeping with these favorable performance characteristics, peCHYRON was used to accurately reconstruct complex cell lineage relationships and event histories in proof-of-concept experiments that exploit the information available in temporally ordered records.

The novel architecture of peCHYRON makes it particularly well-suited to long-term lineage tracing and temporally-resolved multiplexed signal recording in animals, two of the biological grand challenges motivating the development of DNA recorders. Deep lineage tracing on the scale of a whole mammal requires a recorder that can continuously operate throughout development and generate durable records capable of distinguishing among billions of cells[26]. peCHYRON may be such a recorder. It already continuously operates for at least 13 days without decreases in the rate of propagation, and it encodes 3.2 bits of information in each sequentially inserted signature mutation such that a record containing just 11 signature mutations has ~40 billion possible states, similar to the number of cells in an adult mouse. Decoding the complex history of mammalian cell signaling requires a multiplexable recorder that can log a large number of signaling events in order. peCHYRON may be exceptionally well-matched for this task. In our lineage tracing experiments, we showed that at least 41 distinct pegRNAs are available for sequential insertion at the peCHYRON locus during recording (**Supplementary Table 2**). If distinct biological signals are linked to the expression of the pegRNAs through inducible promoters, peCHYRON records will contain the order of multiple signals experienced, reflected in the order of the exact pegRNAs used to propagate each insertion. Applications of peCHYRON beyond the proof-of-concept

experiments shown here will capitalize on these new opportunities in the quest to understand the detailed histories of individual cells in animal biology and development.

Acknowledgements: We thank Christine Duong and Seanjeet K. Paul for technical assistance; members of the Liu Laboratory, Olga Razorenova, and Jordan Woytash for helpful discussions; and David Liu (Broad Institute) and Hana El-Samad (UCSF) for plasmids. This work was funded by NIH grants 1R35GM139513, 1DP2GM119163, and 1R21GM126287 to CCL; NIH grant 1K99GM140254 to TBL; NSF GRFP and AHA Predoctoral Fellowships to CKC; and a fellowship from the NSF-Simons Center for Multiscale Cell Fate Research (NSF Award 1763272) to TBL. VJH is supported by Medical Scientist Training Program grant T32-GM008620. 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.

Author contributions: TBL, CKC, and CCL conceived the project. TBL, CKC, and CCL designed experiments. TBL, CKC, GL, MF, AS, and CADH performed experiments following protocols developed by TBL and CKC. TBL, CKC, VH, and CCL designed and performed analyses. TBL, CKC, and VH wrote code. TBL, CKC, and CCL wrote the paper, with input from all authors. CCL and TBL procured funding and oversaw the project.

Supplementary Material Contents:

Supplementary Tables

Supplementary Table 1. Data and entropy calculations underlying **Figure 3**.

Supplementary Table 2. Data and entropy calculations underlying **Figure 4a-c** and **Extended Data Figure 4**.

Supplementary Table 3. Guide to plasmids used, HTS datasets available at the NCBI Sequence Read Archive, HTS primers, and cell lines.

References

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

2. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Sci New York N Y* 353, aaf7907 (2016).

3. Kalhor, R. *et al.* Developmental barcoding of whole mouse via homing CRISPR. *Sci New York N Y* 361, eaat9804 (2018).

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

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

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

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

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

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

10. Chan, M. M. *et al.* Molecular recording of mammalian embryogenesis. *Nature* 570, 77–82 (2019).

11. Bowling, S. *et al.* An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells. *Cell* (2020) doi:10.1016/j.cell.2020.04.048.

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

13. Chow, K.-H. K. *et al.* Imaging cell lineage with a synthetic digital recording system. *Science* 372, eabb3099 (2021).

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

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

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

17. Frieda, K. L. *et al.* Synthetic recording and in situ readout of lineage information in single cells. *Nature* 541, 107–111 (2017).

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

19. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576, 149–157 (2019).

20. 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).

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

22. Shannon, C. E. Prediction and Entropy of Printed English. *Bell Syst Tech J* 30, 50–64 (1951).

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

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

25. Karacosta, L. G. *et al.* Mapping lung cancer epithelial-mesenchymal transition states and trajectories with single-cell resolution. *Nat Commun* 10, 5587 (2019).

26. Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Ann Hum Biol* 40, 463–471 (2013).

Methods

Plasmid cloning. Plasmids were made by standard Gibson assembly, *in vivo* recombination, or Golden Gate assembly. All plasmids are listed in **Supplementary Table 3** and annotated full plasmid maps are available at github.com/liusynevolab/peCHYRON-plasmids. Pools of PiggyBac cargo plasmids that can be used to make peCHYRON cell lines will be made available from Addgene. All plasmids to be used for transfection were purified with HP GenElute Midi or Mini kits (Sigma # NA0200 and NA0150).

For polymerase chain reactions (PCRs), Q5 Hot Start High-Fidelity DNA Polymerase or Phusion Hot Start Flex DNA Polymerase (New England Biolabs) were used. PCR reagents were purchased from NEB, and all primers were synthesized by Integrated DNA Technologies (IDT).

Plasmids encoding PE2 (Addgene plasmid #132775[19]; http://n2t.net/addgene:132775; RRID:Addgene_132775) and AncBE4Max (Addgene plasmid #112094[27]; http://n2t.net/addgene:112094; RRID:Addgene_112094) were a gift from David Liu (Broad Institute). PiggyBac cargo plasmids used in **Figure 3** and shown in **Extended Data Figure 3** and pegRNA-expressing plasmids used in **Figure 4** and **Extended Data Figure 4** were created using the Mammalian Toolkit (Addgene article #28197510[28]), which was a gift from Hana El-Samad (UCSF).

Cell culture and transfection. HEK293T cells were procured from ATCC (CRL-3216). They were not otherwise authenticated or tested for mycoplasma contamination. All cell culture was performed in DMEM, high glucose, GlutaMAX™ Supplement (Gibco #10566024), supplemented with 10% FBS (Sigma #12306C), at 37°C and 5% CO_2.

HEK293T cells were transfected with Fugene (Promega #E2311), using a ratio of 1 μg DNA to 3 μL Fugene mixed together in serum-free DMEM. In all cases, the target site for mutation was the genomic HEK293 site3.

To create the 293T-peCHYRON$_{6xHis}$ cell lines, HEK293T cells were transfected with plasmids expressing PE2 and a pegRNA that directs the installation of a 20-bp insertion based on the 6xHis sequence, then single colonies were isolated in two rounds of colony picking, dilution of isolated cells, and plating. Integration into the targeted genomic locus and how many of the three copies of this chromosome were modified were verified by deep sequencing. 293T-peCHYRON$_{6xHis\text{-}clone3}$, in which 2 of 3 chromosomes were modified, was used for the experiments in **Figure 2c**; **Extended Data Figure 1a**, **d**, and **e**; and **Extended Data Figure 2a-b**. 293T-peCHYRON$_{6xHis\text{-}clone2}$, in which ~1 of 3 chromosomes were modified, and 293T-peCHYRON$_{6xHis\text{-}clone8}$, in which 2 of 3 chromosomes were modified, were used in **Extended Data Figure 1c**. 293T-peCHYRON$_{6xHis\text{-}clone4}$, in which 1 of 3 chromosomes was modified, was used in **Extended Data Figure 2c**, **Figure 4**, **Extended Data Figure 4**, and **Extended Data Figure 5**.

To create the 293T-peCHYRON$_{PB14v1}$ cell line, 293T-peCHYRON$_{6xHis\text{-}clone4}$ cells were transfected with equal amounts of 14 PiggyBac cargo vectors (**Extended Data Figure 3**, **Supplementary Table 3**) that express PE2 and a total of 42 pegRNAs, along with 1/10 total plasmid mass of Super PiggyBac Transposase Expression Vector (System Bioscience, Inc. #PB210PA-1). Stable integrants were selected with 1-2 μg/mL puromycin.

Screening pegRNAs for efficient prime editing. Unless stated otherwise, experiments toward finding efficient parts for peCHYRON utilized transient transfections of constructs encoding the indicated components with 3 days incubation before ending the experiment by freezing cell

pellets. This pertains to **Figures 2a-c**, **Extended Data Figure 1a**, **Extended Data Figure 1c-e**, and **Extended Data Figure 2a**. For the experiment shown in **Extended Data Figure 2b**, plasmids encoding PE2 and a pegRNA installing each B sequence were transfected into $293T_{6xHis}$ cells in two rounds over 8 days. Then, a plasmid encoding PE2 and libraries encoding the corresponding 6xHis-inserting pegRNAs were transfected into each sample, and cells collected and analyzed after 2 days.

Lineage tracing assay. For the reconstruction shown in **Figure 4a-b** and **Extended Data Figure 4**, 293T-peCHYRON$_{6xHis\text{-}clone3}$ cells in a 6-well dish were transfected with 1.5 μg pCMV-PE2 and 0.5 μg of a mix of equal amounts of 42 pegRNA-expression vectors (**Supplementary Table 3**). These cells were allowed to grow for 8 days, passaging once. Then ("day one"), 20,000 of these cells were plated in each of 4 wells of a 96-well plate, then transfected the next day with the same mix of plasmids. On day four, when they had expanded to approximately 135,000 cells per well, the cells were trypsinized and each well was split into 2 wells of a 24-well dish, before re-transfecting on day five. On day seven, the cells were trypsinized and the entire contents of each well were moved to one well of a 12-well dish. On day nine, when cells had expanded to approximately 750,000 cells per well, each well was split into two wells of a 6-well dish. On day eleven, when each well had expanded to approximately 1.6 million cells, all wells were collected and analyzed by amplicon sequencing.

Amplicon sequencing library preparation. Genomic DNA was isolated with a QIAamp DNA Mini Kit (Qiagen #51304). After DNA extraction, the site 3 region was amplified by PCR. The primers contained partial Illumina adapters and a 5 – 7 nt sample-specific barcode (**Supplementary Table 3**). The PCR reaction was performed with Phusion HotStart Flex DNA Polymerase with GC buffer, with or without DMSO (NEB), and the following protocol: 98°C, 1 min;

(98°C, 10 s; 58°C, 30 s; 72°C, 30 s-1 min.) 5 cycles; (98°C, 10 s; 68°C, 30 s; 72°C, 30 s-1 min.) 25-30 cycles; 72 °C, 5 min.

For the experiments shown in **Figure 2a-b**; **Extended Data Figure 1a**, **c**, and **d**; and **Extended Data Figure 2c**, reactions were pooled and purified by binding to AMPure beads (0.9 beads:1 sample). The libraries were sent to Genewiz, Inc. for Amplicon-EZ sequencing, where they were further amplified to incorporate the TruSeq HT i5 and i7 adaptors and then sequenced on an Illumina HiSeq 2500 with a paired-end 250 protocol.

For the experiments shown in **Figure 2c**, **Extended Data Figure 1e**, **Extended Data Figure 2a-b**, **Figure 3**, **Figure 4**, **Extended Data Figure 4**, and **Extended Data Figure 5**, reactions were purified individually by binding to AMPure beads (0.9 beads:1 sample). Then, they were further amplified to incorporate the TruSeq HT i5 and i7 adaptors, using Q5 High Fidelity DNA Polymerase with GC enhancer, for 15 cycles. The amplified libraries were pooled and purified again by binding to AMPure XP (Beckman-Coulter #A63880) beads (0.9 beads:1 sample). They were subsequently sequenced on an Illumina HiSeq using a paired-end 150-bp protocol at Novogene, Inc.

Deep sequencing analysis. The sequences retrieved by HTS were first demultiplexed based on their barcodes. For the experiments shown in **Extended Data Figure 1e** and **Extended Data Figure 2a-b**, and to detect deletion mutations in **Figure 3c**, data were analyzed as in[14]. Scripts are available at github.com/liusynevolab/CHYRON-NGS. For the experiments shown in **Figure 2a-b**; **Extended Data Figure 1a**, **c**, and **d**; and **Extended Data Figure 2c**, data were analyzed with CRISPResso2[29]. For the high-throughput pegRNA screen shown in **Figure 2c**, 20-bp insertion sequences were extracted from fastq files by simply grabbing the 20-letter sequence at

the expected edit site, then comparing it to the wt (unedited) sequence. If the extracted sequence exceeded a Hamming Distance cutoff from the unedited sequence, it was considered a real insertion for further analysis. All insertion sequences were tabulated, and instances of identical insertions were tallied. To calculate enrichment scores, Illumina sequencing reads from the pegRNA hp-miniprep pool (the plasmids that were transfected into $293T_{6xHis}$ cells at the start of the experiment) were analyzed in the same manner. The abundance of each RT template sequence in the original pegRNA pool was tabulated. Finally, the enrichment of each 20-bp insertion sequence was calculated as follows, where *g* is the proportion of genomic reads bearing that insertion sequence, and *m* is the proportion of library plasmid reads bearing that insertion sequence:

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

(1)

For **Figure 3**, **Figure 4**, and **Extended Data Figure 4**, for each sample, associated forward and reverse reads were merged (PEAR 0.9.10[30]). Insertion sequences were extracted by finding the expected sequence motif upstream of the edit site, then searching for the expected sequence motif downstream of the edit site. For each read, the insertion length was increased in increments of 20 bp until the expected sequence motif downstream of the edit site was found. The full insertion sequence was extracted and the 17-bp propagator sequences in each insertion were compared to the known propagator sequences to ensure only legitimate insertions were grabbed. The 3-bp signature mutations in each insertion were converted to 1-character symbols to improve the ease of interpreting results. Insertion sequences represented by 1-character signature mutations were tabulated, and the counts of each insertion sequence were tallied.

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

Accurate determination of recording efficiency by sequencing. As described above, efficient Illumina amplicon library preparation requires an initial PCR that amplifies the genomic site of interest and adds partial Illumina adapters, followed by purification with 0.9 volume AMPure beads to 1 volume sample, which effectively removes unwanted primer dimers, before a final PCR adds the full adapter sequences. To detect peCHYRON loci with many 20-bp sequential insertions by high-throughput paired-end 150-bp Illumina sequencing, we minimized the size of our amplicon sequencing PCR products to the extent possible, reserving most of the read length for insertions. Accordingly, an unmodified peCHYRON locus would produce an initial PCR product that would be only 134 bp. Owing to its small size, we hypothesized that there might be preferential loss of unmodified peCHYRON loci as well as smaller peCHYRON loci with fewer insertions in the 0.9:1 AMPure bead purification. Loss of such sequences before sequencing would inflate our estimation of peCHYRON's efficiency in generating long sequences through sequential insertion. Therefore, we performed an experiment to ensure that we accounted for this purification bias and avoided overestimating the efficiency of peCHYRON, as follows.

In order to test the purification bias experienced by DNA molecules of different sizes, we first obtained a sample containing peCHYRON loci of various relevant sizes representing 0 to 6 rounds of sequential insertion. This sample was taken from an experiment in which peCHYRON edits accumulated over multiple rounds of transient transfection (as in **Figure 4**). Initial library preparation PCRs were performed using the same method as for the experiments in **Figure 3**, **Figure 4**, and **Extended Data Figure 4**. Three technical replicate PCRs were performed. Then, each replicate PCR was divided into two aliquots to compare the DNA sizes retained by different purification methods. PCRs were purified with the following ratios of AMPure beads to sample: [0.9 volume beads to 1 volume sample], as in our sequencing library preparation protocol, and

[1.8 volume beads to 1 volume sample], which selects for double-stranded DNA of all possible peCHYRON loci sizes (134 bp or larger)[32]. The molar amounts of purified DNA of each size were determined on an Agilent BioAnalyzer using the DNA High Sensitivity kit (#5067-4626). To normalize for the amount loaded in each BioAnalyzer capillary, the amount of DNA at each size was expressed as a percentage of the total moles of DNA detected in that purified sample (**Extended Data Figure 5a**). For DNA sizes corresponding to each of the first four "rounds" of recording (*e.g.*, 134 bp total or 0 bp insertion corresponds to 0 rounds of recording, 154 bp total or 20 bp insertion corresponds to 1 round of recording, etc.), DNA amounts were above the limit of detection in all samples. Therefore, these values were used to calculate a "relative compensation factor" (rcf) to be used to translate molar amounts of DNA of that size observed after 0.9:1 AMPure bead purification (e.g., by sequencing) to the amount that would be detected without purification bias. To determine the rcf for each size, the ratio of DNA detected after purification with [1.8:1 beads to sample] to DNA detected after purification with [0.9:1 beads to sample] was calculated (**Extended Data Figure 5b**). For the DNA sizes corresponding to the first four rounds of recording, the rcf was calculated by averaging the ratios from each of the technical replicates. These values were: 134 bp, rcf = 1.243; 154 bp, rcf = 1.052; 174 bp, rcf = 0.8531; 194 bp, rcf = 0.6351; 214 bp, rcf = 0.1135. To calculate rcf values for sizes corresponding to 5 and 6 rounds of recording, which were readily detectable by sequencing but not by BioAnalyzer analysis, a simple linear regression was performed in Prism on **Extended Data Figure 5b**. Specifically, where *x* is the DNA length in bp, *z* is the abundance observed by sequencing after purification with [0.9 volume beads:1 volume sample], and *a* is the abundance corresponding to purification without bias:

$$rcf = -0.009333x + 2.484$$

(2)

$$a = rcf * z$$

(3)

Using these equations, we calculated that for 234 bp (corresponding to 5 rounds of insertion), rcf = 0.3001 and for 254 bp (corresponding to 6 rounds of insertion), rcf = 0.1134. We consider linear regression a very conservative method to determine rcf values for 5 and 6 rounds, as experiments showing the degree to which smaller DNA is excluded by [0.9 beads:1 sample] purification suggest that a linear extension of compensation values from 134-214 bp is likely to overestimate the extent to which 234-254 bp fragments are enriched[32]. Compensation using these rcf values was performed for all data in **Figure 3**. Namely, the number of reads for each particular size was multiplied by the corresponding rcf before calculating the percent of reads of each particular size, used to determine recording efficiency. In earlier experiments, much larger PCR products were used, so this normalization was not necessary. In later experiments, recording efficiencies were not measured, so this normalization was not relevant.

Calculation of recording accuracy. Sequences of peCHYRON loci were used to determine the extent to which unintended edits are made in each round of peCHYRON recording. Insertion and deletion sequences at the peCHYRON locus, and the abundance of each, were determined and ordered by length. Substitution mutations were not considered in this analysis, as true substitutions were rare. Each indel size was binned to the most-similar “round” of recording. For example, deletion mutations, unedited sequences, and insertion mutations 1-9 bp in length were binned with 0 rounds of recording, and insertions 10-29 bp in length were binned with 1 round of recording. Sequences were considered “correct” if they were the exact expected length (e.g., a 20-bp insertion for one round of recording) and “indels” if they were any other length. The ratio of indels to correct edits was calculated for each round in the sample shown in **Figure 3c-d** and averaged to give a value of 0.98%.

Modeling the expected rate of editing over many rounds. To determine if editing efficiency at the peCHYRON locus remains constant as propagation proceeds, experimental data was compared to a theoretical model wherein efficiency stays constant. For this model, first order rate laws were used to describe the rate of converting peCHYRON loci from having n edits to n+1 edits. Separate rate equations were written for loci bearing different numbers of edits, and two different efficiency values were implemented to reflect the different efficiencies of A->B and B->A pegRNAs. Example rate equations describing the conversion from wt to singly edited loci, singly to doubly edited loci, and doubly to triply edited loci are shown below.

$$\frac{d[wt]}{dt} = -\eta_{A\rightarrow B} * [wt]$$

(4)

$$\frac{d[1\ edit]}{dt} = \eta_{A\rightarrow B} * [wt] - \eta_{B\rightarrow A} * [1\ edit]$$

(5)

$$\frac{d[2\ edits]}{dt} = \eta_{B\rightarrow A} * [1\ edit] - \eta_{A\rightarrow B} * [2\ edits]$$

(6)

where t is in units of days, $\eta_{A\to B}$ and $\eta_{B\to A}$ denote the per-day editing efficiencies of A->B and B->A pegRNAs, respectively, and brackets denote concentrations (*i.e.*, percentage of loci with the specified number of edits).

The constant efficiency values $\eta_{A\to B}$ and $\eta_{B\to A}$ were calculated from the first A->B and B->A edits observed in experimental data from **Figure 3b**. Specifically, equations (4) and (5) above were integrated, and the experimental data provided values for the percentages of loci with wt sequences or single edits at a specific timepoint (t = 13 days). Integration of equation (4) was straightforward, while equation (5) was integrated using Mathematica. The integrated equations are shown below:

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

(7)

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

(8)

Equation (7) was used to solve for $\eta_{A\to B}$ with simple algebra. Equation (8) was solved numerically to find the value of $\eta_{B\to A}$.

Following the same pattern as equations (4)-(6) above, rate equations were written for loci bearing all insertion lengths up to 6 edits. The resulting system of ordinary differential equations (ODEs) was solved in Mathematica using the $\eta_{A\to B}$ and $\eta_{B\to A}$ values calculated from experimental data. This yielded the expected breakdown of insertion lengths at the end of the 13-day timecourse, which was subsequently compared to experimental data to show the match between experimental

data and the constant editing rate model. The Mathematica notebook with this system of ODEs is available at github.com/liusynevolab/peCHYRON.

Lineage reconstruction. To reconstruct cell lineage, we created a list of all insertion sequences in each of the 13 wells used for the analysis. Each insertion has an abundance, based on the number of high-throughput sequencing (HTS) reads that include that exact insertion sequence, and a length, equal to the number of peCHYRON insertions at the recording site. For our initial analysis, the researcher performing the analysis (TBL) was not told which well was which. We refined the list for each well to include only those insertions that that passed two filtering steps: 1) a length filter that excluded any sequences with 3 or fewer inserted sequences and 2) an abundance filter that removed everything after a decreasing convex knee found using the kneedle algorithm[23] (**Extended Data Figure 6**). For each sequence in this set, we generated all possible subsequences containing the first insertion (*i.e.* prefixes) and assigned a weight equal to the inverse of the number of prefixes. This results in a multiset of prefixes for each sequence. Afterwards, the distance between the multisets were found using the generalized Jaccard distance for multisets. We used this set of distances to reconstruct the relationships using the UPGMA[31] hierarchical clustering algorithm. (github.com/scipy/scipy/blob/v1.2.1/scipy/cluster/hierarchy.py#L411-L490). The two algorithms (Jaccard and Prefix Jaccard) were compared by computing the Robinson Foulds score[33], a metric that compares a reconstruction to a ground-truth tree, for each reconstruction (**Extended Data Figure 4**). For each percentage of downsampling between 99 and 5%, 1000 different data downsampling operations were performed. Robinson Foulds scores were computed for each using the ete3 package[34] (http://etetoolkit.org/).

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

Event reconstruction. To make heatmaps of signature enrichment at each position in the peCHYRON recording locus, every read with exactly 4 insertions was isolated. For every possible signature, the fraction of insertions possessing that signature was calculated at each insertion position 1-4, and a table containing this information was output from the analysis pipeline. The table was converted to a heatmap, using conditional formatting in Excel.

Entropy calculations. To calculate Shannon entropy[21], we first made a table of all the signature sequences in the relevant dataset with, for each sequence, the number of times it was observed (the "count"). We did one calculation for signatures inserted at "odd" positions (1, 3, 5, etc.) and one for "even" signatures. Once a table of signatures and counts (c) was created, we calculated the proportion (*p*) for each sequence (equation shown here for sequence i).

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

(9)

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

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

(10)

All analyses were done in Excel (**Supplementary Tables 1-2**).

Statistical analyses. In all cases, biological replicates were derived from different populations of cells that were manipulated separately throughout the experiment. For technical replicates, cells were grown and manipulated, and DNA extracted, together. All procedures downstream of DNA extraction were performed separately.

Figure preparation. The **Graphical Abstract, Figure 1, Extended Data Figure 1b, Figure 4d,** and a portion of **Figure 4a** were prepared with InkScape. **Figures 2c and 4e** were plotted in Excel, then converted to scalable vector graphics with InkScape. A portion of **Figure 4a** was prepared at biorender.com. The plots in **Figure 4b-c** and **Extended Data Figure 4** were generated using the hierarchy.dendrogram function in matplotlib (scipy.org).

Data Availability Statement

All scripts created for this manuscript are available at github.com/liusynevolab/peCHYRON. All NGS data sets will be deposited at the NCBI's Sequence Read Archive upon paper acceptance. Full plasmid maps are available at github.com/liusynevolab/peCHYRON-plasmids. Plasmids necessary to carry out peCHYRON lineage tracing will be available at Addgene. See **Supplementary Table 3** for a guide to these data and reagents. Please contact CCL and TBL for cell lines.

References for Methods

27. Koblan, L. W. *et al.* Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat Biotechnol* 36, 843–846 (2018).

28. Fonseca, J. P. *et al.* A Toolkit for Rapid Modular Construction of Biological Circuits in Mammalian Cells. *Acs Synth Biol* 8, 2593–2606 (2019).

29. Clement, K. *et al.* CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat Biotechnol* 37, 224–226 (2019).

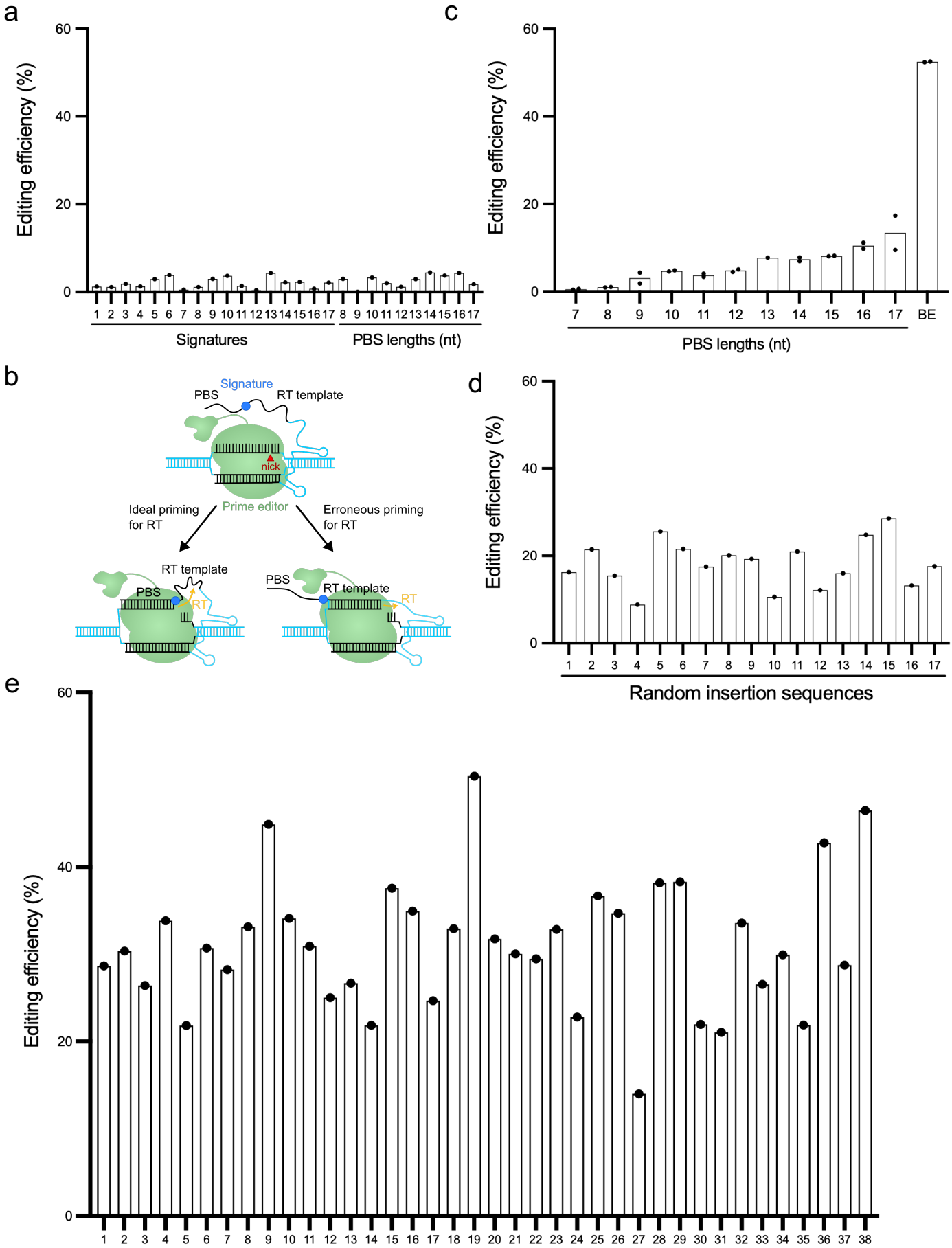
30. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinform Oxf Engl* 30, 614–20 (2013).

31. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. *Arxiv* (2011).

32. Bronner, I. F., Quail, M. A., Turner, D. J. & Swerdlow, H. Improved Protocols for Illumina Sequencing. *Curr Protoc Hum Genetics* 79, 18.2.1-18.2.42 (2013).

33. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math Biosci* 53, 131–147 (1981).

34. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* 33, 1635–1638 (2016).



Extended Data Figure 1. Testing alternative pegRNA candidates for propagation.

Extended Data Figure 1. Testing alternative pegRNA candidates for propagation. a, Efficiency of iterative editing with a pegRNA that can both target and insert the same 20-bp sequence (A->A) was low. The 20-bp sequence was derived from the 6xHis sequence. Constructs expressing PE2 and appropriate pegRNAs were transfected into $293T_{6xHis}$ cells. The signature mutations were randomized, and different PBS lengths were scanned. **b,** Proposed mechanism of failed editing with a pegRNA that both targets and inserts the same 20-bp sequence (A->A). The PBS and RT template sequence share identity, so priming can occur erroneously in the RT template instead of the PBS, yielding no net insertion. **c,** Although the 6xHis sequence readily recruited Cas9, the efficiency of inserting the wt site3 sequence at a 6xHis target site was low. Here, site3 was the original sequence at the recording locus and 6xHis had been inserted (A->B) to make the $293T_{6xHis}$ cells used in this experiment. Thus, to achieve propagation, the efficiency of adding back site3 was tested (B->A). For optimization, multiple lengths of the PBS of the site3-inserting pegRNA were tested. Each sample is a library of all possible signatures with the indicated PBS length. To ensure that the 6xHis sequence could be efficiently targeted by a pegRNA-directed Cas9, plasmids encoding the C->T base editor AncBE4Max and a pegRNA that targets 6xHis were co-transfected into $293T_{6xHis}$ cells, and the efficiency of a base edit in the 6xHis sequence was measured. Points represent biological replicates. **d,** Insertion efficiencies of random pegRNAs that target 6xHis were variable and low. In this setup, 6xHis is sequence A and the 20 random insertion sequences are B (A->B). **e,** Rare sequences that can efficiently insert downstream of 6xHis were identified. The top hits from the high-throughput screen shown in Figure 2c were individually cloned, and their insertion efficiencies were assayed. All experiments were conducted in $293T_{6xHis}$ cells with PE2. Insertion sequences were ordered from left to right based on their enrichment scores in the original screen.

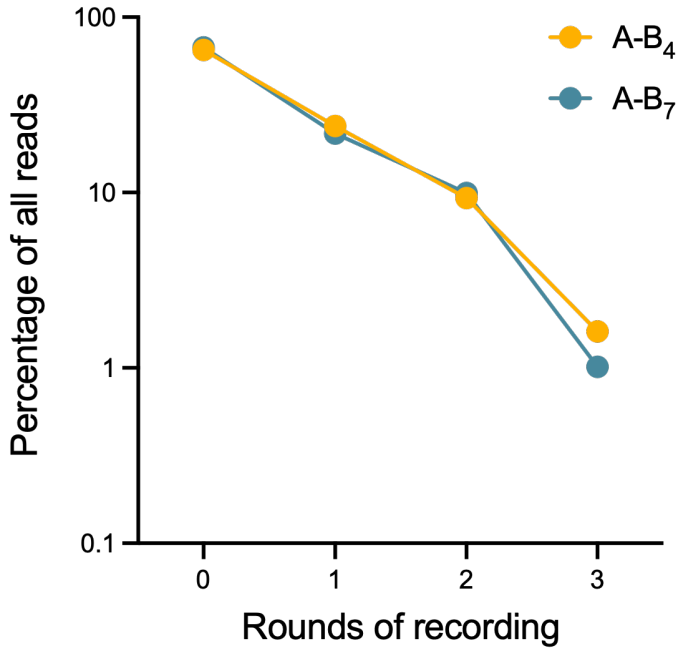
a

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

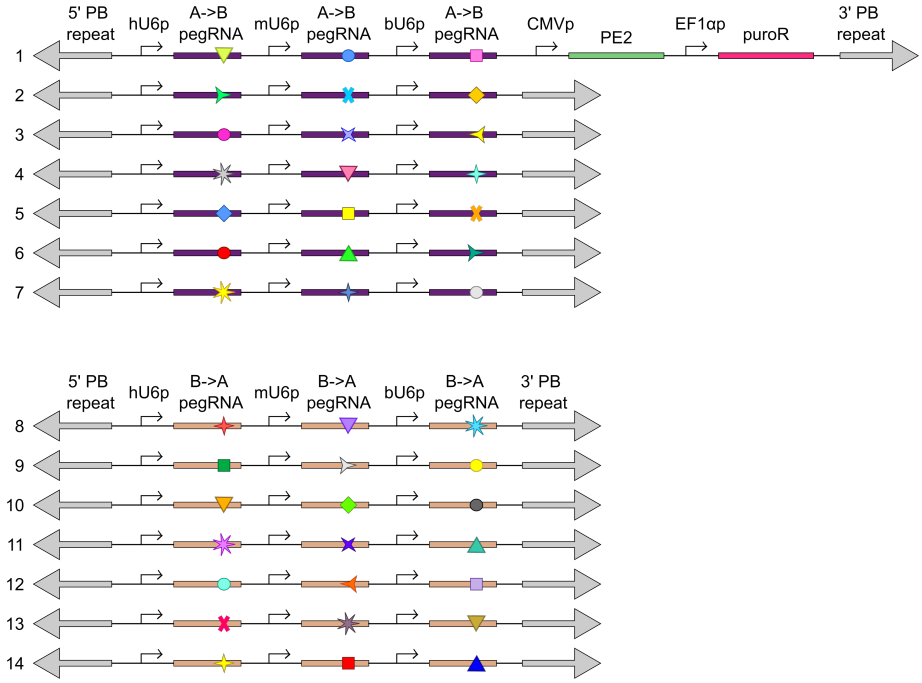
b

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

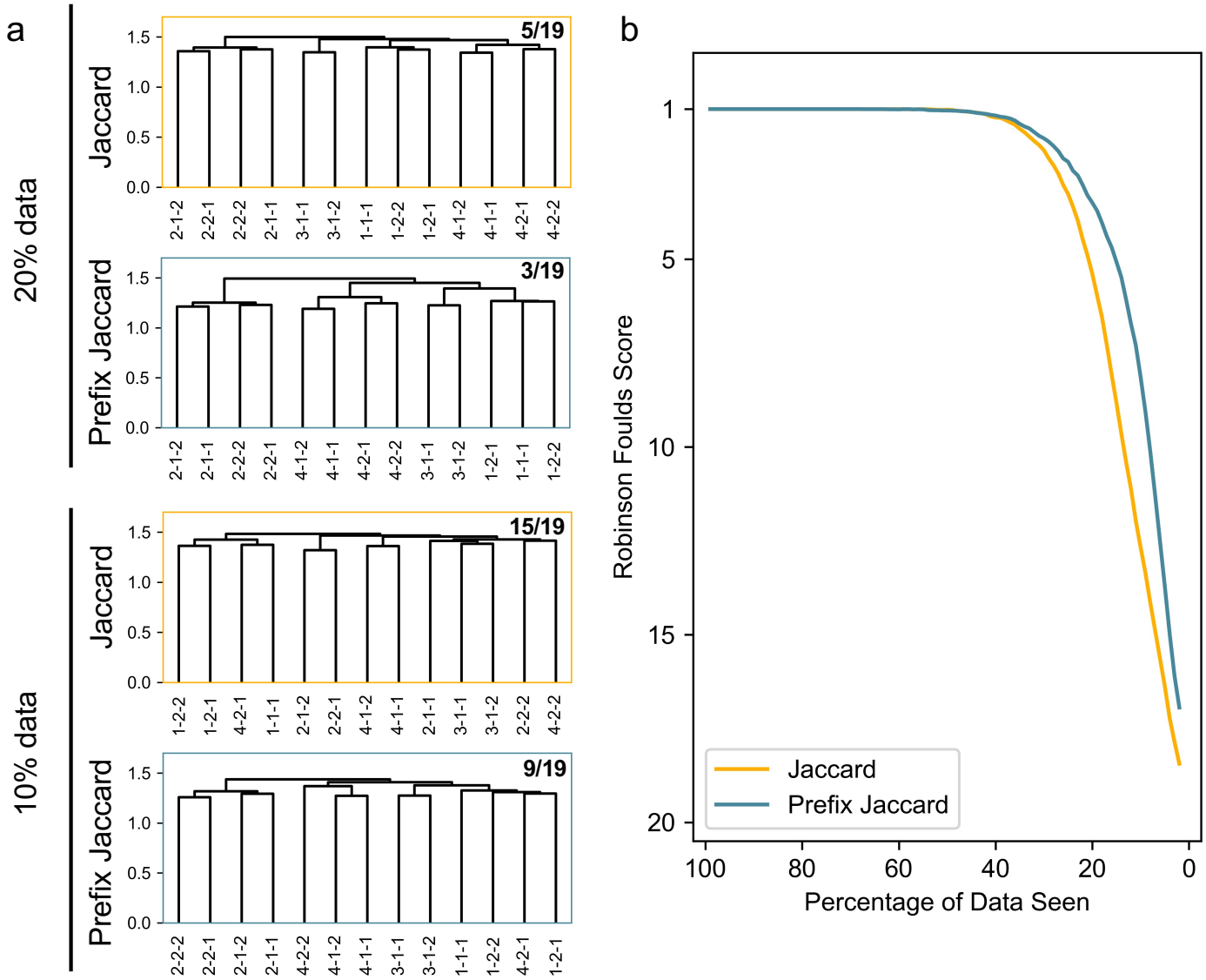
c



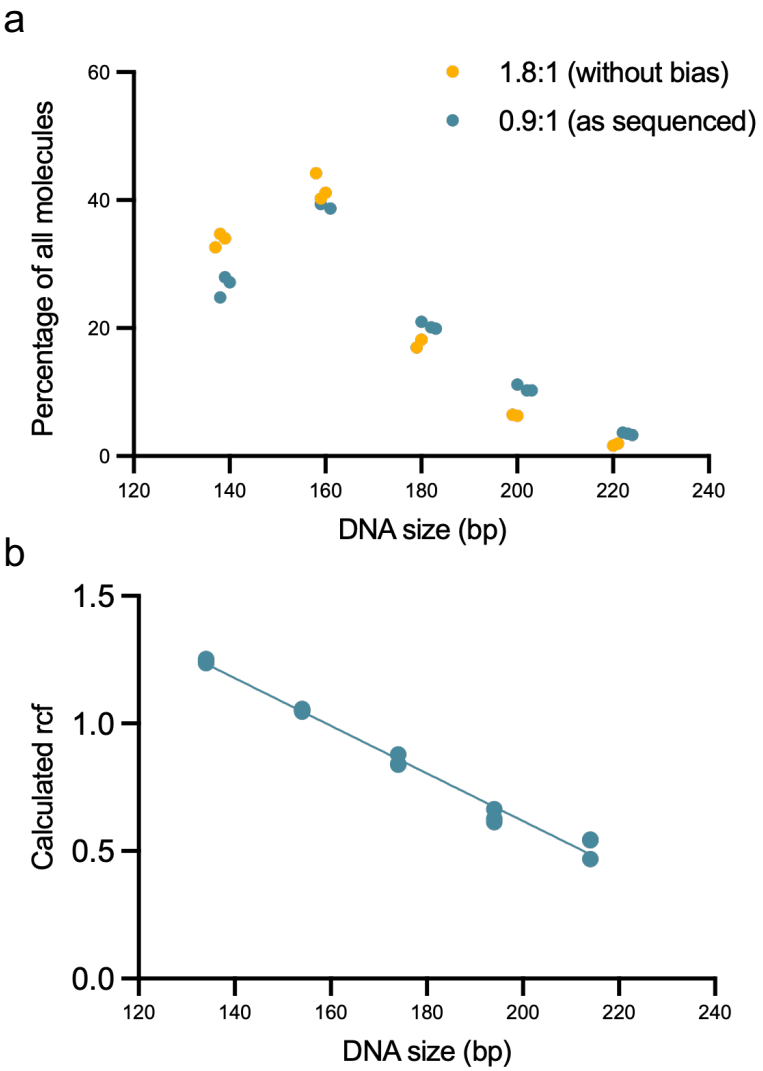
Extended Data Figure 2. Identification of efficient propagating pegRNAs. a, Heatmap of editing efficiencies for the top ten A->B propagator sequences identified in Extended Data Figure 1e (B_1-B_{10}). PBS lengths ranging from 9-15 nts (row labels) were tested with a library of signatures for each PBS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



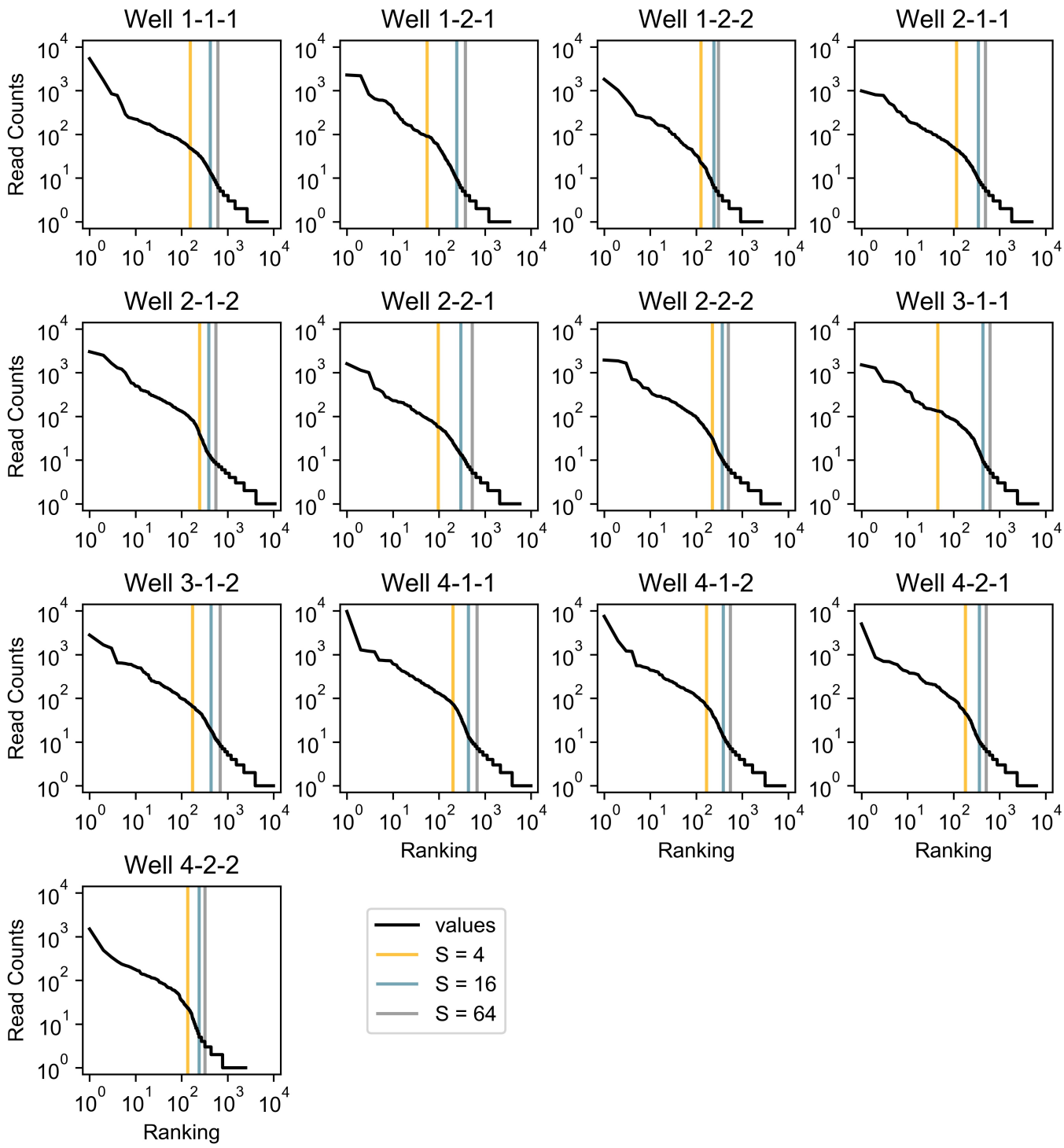
Extended Data Figure 3. Constructs encoding peCHYRON components. Depiction of the 14 PiggyBAC Transposase cargo vectors stably integrated into the genome for the experiments shown in Figure 3. 5' PB and 3' PB repeat are the inverted terminal repeats for PiggyBac transposition. Colorful shapes are unique signature mutations encoded in each pegRNA.



Extended Data Figure 4. Comparison of lineage reconstruction methods. a, Reconstruction of culture splitting patterns from a cell lineage tracing experiment with random downsampling was more accurate when our prefix Jaccard similarity index was used. For each reconstruction, we calculated the Robinson Foulds score, which is the number of changes required to transform the reconstruction into the ground truth tree. The score for each reconstruction is shown in the upper right, divided by the highest possible score (19). **b,** Average Robinson Foulds scores for reconstructions using our prefix algorithm (blue) and the Jaccard distance (orange). This was calculated by downsampling the data to a certain percentage (x-axis) and performing a reconstruction which was then compared to the true tree. Each point was obtained by averaging over 1000 repeats. Note, perfect reconstruction in our case returns a score of 1 because we compare a rooted constructed tree to an unrooted ground truth tree.



Extended Data Figure 5. A method for accurate determination of recording efficiency by sequencing. a, Determination of DNA size bias introduced by our library preparation protocol. peCHYRON loci were amplified from genomes of cells that had undergone 21 days of transient transfection with constructs encoding PE2 and a pair of propagating pegRNAs. Three technical replicate PCRs were performed. Each PCR reaction was then split into two aliquots. One was purified with 0.9 volumes of AMPure beads to 1 volume of sample, as in our standard sequencing library preparation protocol. The other was purified with 1.8 volumes of beads to 1 volume sample, which retains all relevant sizes of DNA. The purified PCR products were assayed on a BioAnalyzer to determine the number of molecules of each size. For each purified sample, each value was normalized to the total DNA molecules detected in that sample. **b**, A relative compensation factor (rcf) for calculating abundance of DNA molecules of each size. For each technical replicate shown in (**a**), the value obtained with the unbiased 1.8:1 purification method was divided by the value obtained with the 0.9:1 purification method. The x-axis shows the size of the amplification products from the peCHYRON locus where 134 bp corresponds to no insertions, 154 bp corresponds to the outcome of 1 round of sequential insertion, 174 bp corresponds to the outcome of 2 rounds of sequential insertion, 194 bp corresponds to the outcome of 3 rounds of sequential insertion, and 214 bp corresponds to the outcome of 4 rounds of sequential insertion.



Extended Data Figure 6. Determination of sequence abundance cutoffs for lineage reconstruction. A log-log plot of ranked read counts for each unique sequence. The kneedle algorithm was used to find a decreasing convex region of maximum curvature. This represented an inflection point where reads of lower counts were likely sequencing artifacts. A sensitivity value of 16 (light blue) was selected for all further downstream calculations. Therefore, 4571 sequences, representing approximately 75% of raw reads, were used for analysis.