Supplementary Information for:

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, Guohao Liang[1], Michelle Ficht[1], Beide Liu[1], Xiaohui Xie[3], Chang C. Liu[1,2,4,5,6] *

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

*Correspondence to ccl@uci.edu

Supplementary Note

Interpretation of TdT's reduced activity when fused to Cas9

We were initially surprised when we could not improve TdT activity by fusing it to Cas9 through a variety of flexible linkers. We observed TdT activity, but it was reduced relative to constructs in which TdT was expressed without any linkage (**Supplementary Figure 2c**). Thus, TdT is inhibited, whether sterically or through misfolding, by fusion to Cas9. It may have some activity in the context of a fusion protein, or the activity we see may be due to cleavage of the flexible linker and liberation of TdT (**Supplementary Figure 2d**). Therefore, we used TdT as an unfused protein throughout this work.

Interpretation of non-root CHYRON sequences observed in the absence of Cas9 and TdT

The non-zero entropy observed at time zero in **Extended Data Figure 4c**, for example, is likely due to a low but significant rate of barcode hopping in our NGS library prep. (For these initial experiments with $CHYRON_{20}$, but not thereafter, the final adaptor incorporation was performed on pooled samples.) The Shannon entropy we observed at the final timepoint for the hgRNA with Cas9 alone, 5 bits, was strikingly similar to the value reported by Kalhor *et al.*, 2017[4] for an hgRNA of similar initial length.

Considerations for calculating Shannon entropy of CHYRON recording

We consider 1.74 bits/bp (**Extended Data Figure 6a**), the figure calculated from the set of all 4-nt sequences observed in our large, deeply sequenced lineage tracing experiment, to be a conservative estimate of the information-encoding capacity of CHYRON, for three reasons: 1) all possible 4-nt sequences were present in our dataset, but some 5-nt sequences were absent (see github.com/liusynevolab/CHYRON-insertion-entropy-calc), suggesting that our estimate is beginning to be limited by sequencing depth, 2) four is the size of the DNA alphabet[25], and 3) TdT engages with

approximately three nts of the primer while adding an additional nt[53], raising the possibility that the preceding three nts could affect which 4th nt is added.

Assessment of the performance of CHYRON$_{17}$, delivered by lentivirus, in primary cells

The final average length of the CHYRON$_{17}$ loci was 5.14 inserted bps (**Extended Data Figure 7c**). Given our observation that TdT inserts, on average, 3.04 bp per round at this protospacer sequence in this primary cell line (**Extended Data Figure 2b**), CHYRON$_{17}$ is likely accumulating insertions in multiple rounds. However, the final average length was lower than we expected, even lower than the 5.7 inserted bps we observed for CHYRON$_{20i}$ (**Figure 4b**). We conclude that CHYRON can function in primary cells, but the exact genomic context affects activity. It is unlikely that the central functions of CHYRON, cutting by Cas9 and writing by TdT, are fundamentally different in primary cells, given that single round insertions in these primary cells are identical to those generated in 293T cells (**Extended Data Figure 2**). In the context of an integrated lentivirus, the hgRNA or protein components of CHYRON may be silenced over time. Alternatively, the DNA sequence encoding the hgRNA within the integrated lentivirus may be less-efficiently targeted by Cas9 and TdT as the experiment proceeds. In the future, these issues may be addressed by adding insulators to or otherwise optimizing the lentiviral construct. Alternatively, mouse embryonic stem or human induced pluripotent stem cells can be precisely genetically engineered to express CHYRON components at safe harbor loci, flanked by insulators, or both.

Theory behind successful lineage reconstruction

In the experiment shown in **Figure 5**, the approximately three doublings between splits ensured that enough cells could acquire an insertion and then divide between splits, and the grow-out at the end of the experiment ensured that our recovery of unique sequences was high. Why is this efficient sampling important? Let us consider a well (Well A), its closest relative (Well B), and a totally unrelated well (Well X). Suppose we detect a specific CHYRON sequence (or a shared substring of the CHYRON sequence

given the ordered nature of insertions in CHYRON) in only two of these three wells. What is the chance that the two wells are Well A and Well B? If the sequence is long, then the chance is high, since a long CHYRON sequence is unlikely to arise independently. However, if the sequence is short, then it is possible that the two wells sharing the sequence are, for example, Well A and Well X, and that the sequence was generated independently in both wells. The critical issue, however, is that the sequence could also be absent in Well B, because 1) the sequence was generated after Well A and Well B split from their common ancestor or 2) the sequence was present also in Well B but was not detected due to sampling inefficiencies. In that case, the short sequence will preferentially assign relatedness to the unrelated wells over the related wells, reducing the accuracy of reconstruction. These considerations are treated further below. We observe below that potentially informative sequences, those that were generated before Well A and Well B split, and are successfully sampled from each well, decline with the square of the proportion sampled. However, potentially misleading sequences, that were generated independently in, for example, Well A and Well X, decline only linearly with decreased sampling. Therefore, one must use sufficiently long CHYRON sequences and ensure sufficient sampling of the sequences in all wells, a conclusion generalizable to all lineage reconstruction studies from self-mutating DNA recording systems. Longer, more information-dense sequences can partially, but not completely, compensate for poor sampling.

In early efforts, we attempted to reconstruct relationships among populations of cells using growth, library prep, and sequencing protocols that captured an insufficient percentage of the cells in each population. These attempts were unsuccessful because sequencing a high proportion of each population is essential to successful lineage reconstruction. The CHYRON locus from a cell can potentially report on the relationships between the populations in which the cell ends up only if the cell acquires an insertion, then the cell divides one or more times, and then the descendants are split so at least one descendant is distributed to each daughter population. However, the fraction of these potential “reporter” CHYRON loci that will actually give useful information declines with the square of the fraction of each population sampled. Specifically, if r = number of insertions in a data set that arose, then the cell divided, then the

daughter cells were split into separate wells and p = proportion of cells in each population that are sampled, and b = informative loci.

$$b = rp^2 \quad (3)$$

For example, consider the following scenario: if only 10% of each well is sampled, the percentage of potential reporter loci that can report accurately is only 10% squared, or 1%. If only a small fraction of each population is sampled, the chance of sampling related loci from both of two or more related populations will be very low.

As we attempt to reconstruct population lineage correctly, we must also consider potentially misleading sequences. “Homoplasies,” or insertions with the same sequence that were generated independently, could cause the unrelated wells in which they arose to be incorrectly considered related. The misleading effect of homoplasy can be minimized by limiting reconstruction to insertions that are long enough that homoplasies are unlikely. Assign the following additional variables:

H = average Shannon entropy for insertions of a given length

n = the average number of insertions of that length in a well or population

x = the expected number of homoplasic insertions in another well

For an insertion of a given length in well A, the formula for the expected number of homoplasic insertions in other wells (*i.e.* the expected number of homoplasic insertions of that length between one of those populations and an unrelated population) is

$$x = pn\left[1 - \left(\frac{2^H - 1}{2^H}\right)^{pn}\right] \quad (4)$$

We may define the variable f (for “effectiveness”) as the expected number of matched insertions between two identical populations (*i.e.*, true matches indicating a lineage relationship) divided by the expected

number of homoplasic insertions between one of those populations and an unrelated population. Combining equations (3) and (4) gives

$$f = \frac{rp^2}{pn\left[1 - \left(\frac{2^H - 1}{2^H}\right)^{pn}\right]} \quad (5)$$

This can be rewritten as:

$$f = \frac{rp}{n\left[1 - \left(\frac{2^H - 1}{2^H}\right)^{pn}\right]} \quad (5)$$

Thus, more information-encoding capacity is required as population relatedness and sampling declines, or the population size increases.

Analysis underlying successful lineage reconstruction

The CHYRON lineage tracing dataset shown in **Figure 5** could be used for an accurate reconstruction of all relationships between wells with no length or abundance cutoffs applied at all (**Supplementary Figure 4a**), but the distances between closely related wells were greatly exaggerated. For more accurate reconstructions, setting an abundance cutoff properly is essential, to remove artifactual “sequences” that result from library prep or sequencing errors, without filtering out too many genuine sequences. For the experiment shown in **Figure 5**, for each percentage abundance value, we plotted the counts of sequences that had that value. This plot should have two peaks, one at the minimum abundance for genuine sequences and one at a lower abundance that represents the minimum abundance of artifactual sequences. We could not use the plot generated for the experiment shown in **Figure 5** to determine a proper abundance cutoff because it had only one peak. There are two possible contributors to this

observation: 1) we did not sequence our library deeply enough (we analyzed ~3M reads per well and each well contained ~1.2M cells with non-root CHYRON loci that should escape PmlI degradation) and/or 2) our library prep process was flawed, so that many sequences that arose as errors during library prep were better-represented in our data set than genuine sequences. To set an abundance cutoff we used a smaller data set that was produced using the same library prep protocol but was fully sequenced and produced two peaks (**Supplementary Figure 4b**). (This data set is from a previous attempt at lineage reconstruction that was very poorly sampled due to cell death and having no grow-out step. The raw data are available in the Sequence Read Archive, and samples are listed in **Supplementary Table 4**.) Thus, we set the abundance cutoff so that all sequences that were observed in at least 0.0139% of all non-deletion reads in a well were included in the analysis of that well.

In our experiment, we found that >40% of long insertion sequences (unlikely to be identical by chance) were identical between sibling wells (**Supplementary Table 2**), suggesting that our sampling efficiency is high. Therefore, it is not surprising that our reconstruction is robust: compared to the ideal reconstruction that results from using insertions 8-15 bp in length (**Figure 5b**), reconstructions using shorter insertions that could be homoplasies, or those restricted to longer, less abundant insertions, only slightly exaggerate the distance between some sibling wells (**Supplementary Figure 4c**).

By artificially degrading our data set, we were able to systematically test the effect of insertion length cutoffs on reconstruction quality. If a large number of artifactual sequences are included in the data set, because the abundance cutoff is set too low, only reconstructions that rely on longer insertions are perfectly accurate (**Supplementary Figure 4d**). When insertions were computationally removed at random from each well to simulate a sampling efficiency that is 25% of the actual experiment, there is a trend toward better reconstruction with insertion length cutoffs of 8-15 bp or 9-15 bp (**Supplementary Figure 4e**). Both of these results can be understood as a reflection of the competition between identical insertion sequences that share a common ancestor and identical insertion sequences that arose independently. The former decline in abundance with the square of the sampling efficiency, as discussed

above, so when sampling is reduced to 25%, identical insertion sequences that share a common ancestor are reduced to 6.25% of their former abundance in the pool of insertion sequences used to generate the reconstruction. In contrast, identical insertion sequences that arise independently are only reduced in abundance linearly with declining sampling efficiency, so 25% of these sequences are still present. Longer identical insertions are less likely to arise independently, so reconstructions that rely on longer insertions are more successful when sampling is limited (until the length cutoff is pushed so high that not enough insertions are available to allow reconstruction). Reducing the abundance cutoff introduces artifactual sequences to the data set. These artifactual sequences can be coincidentally identical to sequences in other wells, but of course they cannot reflect true relationships between wells, since they did not arise during the growth of the cells. Therefore, when the abundance cutoff is set too low, setting a longer length cutoff can promote accurate reconstruction by screening out homoplasic sequences that are coincidentally identical to insertion sequences in unrelated wells. Although it is possible to sample a high proportion of cells in a real-world tissue[16], sampling at high efficiency may be challenging in some applications. The high diversity of sequences that can be recorded at the $CHYRON_{16i}$ locus may therefore be crucial for accurate lineage reconstruction.

Lineage reconstruction in a cell-limiting context in primary cells

As described in **Results** and **Methods**, we performed a cell splitting experiment and lineage reconstruction using primary dermal fibroblasts in culture (**Extended Data Figure 8a**). After infection with lentiviruses carrying $CHYRON_{17}$ (as in **Extended Data Figure 7a**), in two wells of a 24-well plate, cells were allowed to recover for three days before being split into the starting wells. At this point, the cell number had not grown substantially – it was 38,000 at infection and 44,000 after 3 days of recovery – and very few insertions had accrued at the CHYRON locus. After 22,000 cells were plated in each of the starting wells, the cells were allowed to grow for four days, or 1.57 population doublings. Then the 65,000 cells in each well were split into two wells each. Each of these populations was allowed to grow for 14 days, or 2.56 population doublings. Then each of these populations was split into two wells and allowed

to grow for 4 days, or approximately one population doubling. We attempted to use the sequences of the CHYRON loci in the final wells to reconstruct the splitting procedure. For each split, the period of cell growth and insertion accumulation just prior to that split is most important for reconstructing that split. The final split was reconstructed completely accurately, even when up to 60% of the unique insertions in each well were computationally removed (**Supplementary Figure 5a**). Thus, $CHYRON_{17}$ can easily reconstruct lineage when cells have 14 days to accumulate insertions at the CHYRON locus and the population size doubles more than twice. The penultimate split, in which four wells were split into eight, was reconstructed accurately for only two of the four populations. These two successful reconstructions are very likely to be genuine, rather than the result of chance, given that they were maintained when 20% of unique insertions were removed at random in ten replicates (**Supplementary Figure 5a**). Therefore, we conclude that reconstruction after four days of accumulating insertion mutations and less than two population doublings are on the edge of $CHYRON_{17}$'s capabilities. Because $CHYRON_{17}$ accumulated an average of only 5.14 inserted bps (**Extended Data Figure 7c**), it was not surprising that reconstructions were only marginally improved by ignoring short insertions (**Supplementary Figure 5b**). For future use of CHYRON in untransfectable cell types or contexts, it will be critical to increase the information-encoding capacity above 5.14 inserted bps. This could be accomplished by improving the genomic insulation of the $CHYRON_{17}$ locus, or by using a version of $CHYRON_{16i}$, perhaps with multiple copies inserted into the genome of each cell to compensate for the lower initial activity of $CHYRON_{16i}$.

Comparison to a deletion-based recorder

As described in **Results** and **Methods**, we compared lineage reconstruction with our large dataset (**Figure 5**) to lineage reconstruction with the same dataset, transformed by truncation to match the information-encoding characteristics of an hgRNA with Cas9 alone[11]. We chose to compare CHYRON to the hgRNA with Cas9 alone because it is the current technology that encodes the most information per single site. Recording at a single site has several advantages: it is easy to introduce a single recording locus into a new cell[17] and having only one site at a locus avoids the significant problem of simultaneous

double cuts and loss of the intervening locus[2,17], or a requirement for very long loci if sites record small amounts of information and many sites are required[8]. There are currently 3 ways to record at a single site: 1) use a base editor to edit a target[8], to record 1-2 bits; 2) target Cas9 to a site, usually part of an array, to cause primarily deletion mutations in one round, recording up to 4.42 bits[6]; or 3) target Cas9 to an hgRNA to cause primarily deletion mutations in multiple rounds, recording up to 7.97 bits[11].

In our simulated comparison, when sampling was limited, we found that fewer than half of relationships between wells were reconstructed from the truncated dataset, while almost all relationships were reconstructed from the equivalent full-length dataset (**Extended Data Figure 9a-b**). As expected (see **Supplementary Figure 4e**), the best reconstructions from the sampling-limited full-length dataset were achieved when only long insertions were considered (**Extended Data Figure 9c**). The best reconstructions from the sampling-limited, truncated dataset were achieved when all insertions were considered (**Extended Data Figure 9c**).

Prospects for the use of CHYRON in developing organisms

The first requirement for an *in vivo* DNA recording technology is that the technology itself not significantly distort the cellular or developmental processes it aims to record. We showed here that CHYRON activity is compatible with cell proliferation and survival over at least nine days in 293T cells (**Figures 4-5**), and it is extremely likely that CHYRON will be compatible with normal development in a whole mouse as well. Kalhor *et al.*[11] have shown that integration and expression of hgRNAs and Cas9 are compatible with mouse development. Thus, TdT is the only CHYRON component that has not yet been used as part of a DNA recorder in an organism. TdT is a mammalian protein whose expression is normally restricted to a small window during B cell development[54]; in CHYRON, TdT is constitutively expressed and recruited to the Cas9 cut site through its natural interaction with the DSB repair machinery. Therefore, we must exclude the possibility that normal development will be perturbed by TdT acting on endogenous DSBs.

Previous literature suggests that this perturbation will be minimal, at least on a timescale of weeks to months. Bentolila *et al*.[54] created a mouse line in which TdT is expressed as a transgene, not just in early T and pro-B cells post-birth, as it is in wild-type mice, but throughout B-cell development, including in fetal tissue and mature B cells. TdT expression was high enough to perform its normal function of diversifying junction sequences at immunoglobulin loci outside the normal times at which these diversification events usually occur, suggesting the expression level of the TdT transgene is similar to what would be needed for CHYRON recording. No defects were observed. Several types of cancer cells have also been found to express TdT constitutively[55,56], including the non-B-cell-derived Merkel cell carcinoma[57]. This observation may suggest a deleterious consequence of TdT expression in the very long term, but for the time periods more likely to be relevant for lineage tracing in model organisms, it suggests that CHYRON could be used to study carcinogenesis and metastasis in models of cancer. Therefore, at minimum, CHYRON will be compatible with recording of B-cell and cancer development *in vivo*.

DNA recorders must also be capable of recording enough information to accomplish the goals of the recording experiment. Several current DNA recorders have been used to report accurately on known cell lineages during development *in vivo*, although they have not yet reached accurate cell-resolution lineage reconstruction for most tissues[58]. We will consider two technologies that have been successfully used in mice, in order to understand how CHYRON might contribute to lineage tracing during development. Kalhor *et al.*, 2018[11] were able to correctly trace the relationships between six populations of cells in mouse embryos at E14.5, using the data from only one hgRNA (hgRNA #36, see Figure S9 in Kalhor *et al.*) with a maximum Shannon entropy of 7.97 bits (**Supplementary Table 3**). More complex lineage relationships could be traced by taking into account more hgRNAs[11]. Chan *et al.*, 2019[16] were able to reconstruct a cell-resolution lineage tree of a 104,400-cell mouse embryo. For this reconstruction, they used 1,753 unique recorder alleles captured from 15,963 cells. Given that $CHYRON_{16i}$ insertions have a Shannon entropy of 14.6 bits, suggesting that at least ~25,000 unique alleles should be expected from a

sufficiently large experiment, certainly more complex lineages could be traced by CHYRON. Importantly, the greatly reduced risk of homoplasy means that much larger population sizes can be used. One weakness of $CHYRON_{16i}$ is that only 26% of cells acquire insertions after 9 days (**Figure 4c**), compared to the nearly 100% of highly active, longer hgRNAs that are mutated over a similar timeline[11]. This characteristic means that a single CHYRON recorder per cell is likely a better option for larger population sizes, where tracing the lineages of a fraction of the cells is acceptable, and avoiding homoplasy is paramount. To trace lineage at truly single-cell resolution, if even four $CHYRON_{16i}$ loci were integrated in each cell, after nine days we would predict that 42% of cells would have exactly one locus bearing an insertion; 22% of cells would have two loci bearing an insertion, for an expected Shannon entropy of 29.2 bits; 6% of cells would have 3 or 4 loci bearing insertions; and only 30% of cells would have no insertions. CHYRON's increased information-encoding capacity per site, especially when combined with other innovations such as the integration of multiple hgRNAs and reading out by single-cell RNA-seq, will allow more accurate lineage reconstructions on larger populations of cells.

Supplementary References

53. Delarue, M. et al. Crystal structures of a template-independent DNA polymerase: murine terminal deoxynucleotidyltransferase. Embo J 21, 427–439 (2002).

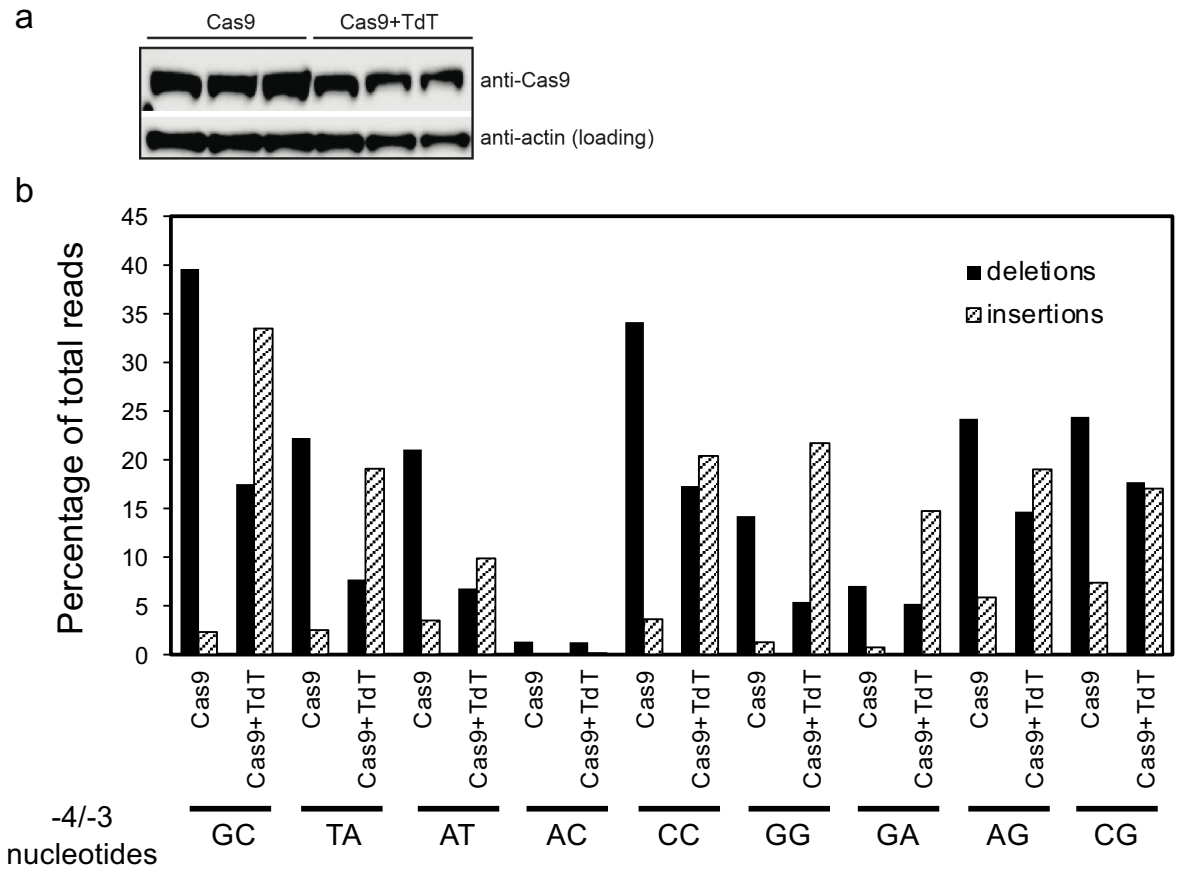
54. Bentolila, L. A. et al. Constitutive expression of terminal deoxynucleotidyl transferase in transgenic mice is sufficient for N region diversity to occur at any Ig locus throughout B cell differentiation. J Immunol Baltim Md 1950 158, 715–23 (1997).

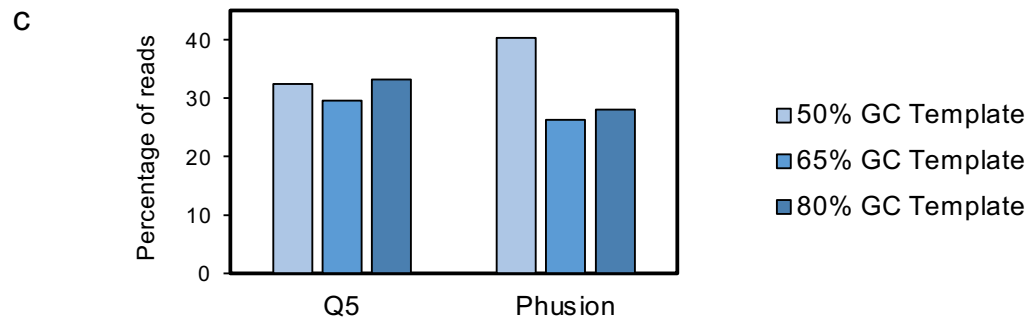
55. Drexler, H. G., Messmore, H. L., Menon, M. & Minowada, J. A Case of TdT-Positive B-Cell Acute Lymphoblastic Leukemia. Am J Clin Pathol 85, 735–738 (1986).

56. Michiels, J. J. et al. TdT positive B-cell acute lymphoblastic leukaemia (B-ALL) without Burkitt characteristics. Brit J Haematol 68, 423–426 (1988).

57. Buresh, C. J., Oliai, B. R. & Miller, R. T. Reactivity With TdT in Merkel Cell Carcinoma: A Potential Diagnostic Pitfall. Am J Clin Pathol 129, 894–898 (2008).

58. Salvador-Martínez, I., Grillo, M., Averof, M. & Telford, M. J. Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? Elife 8, e40292 (2019).

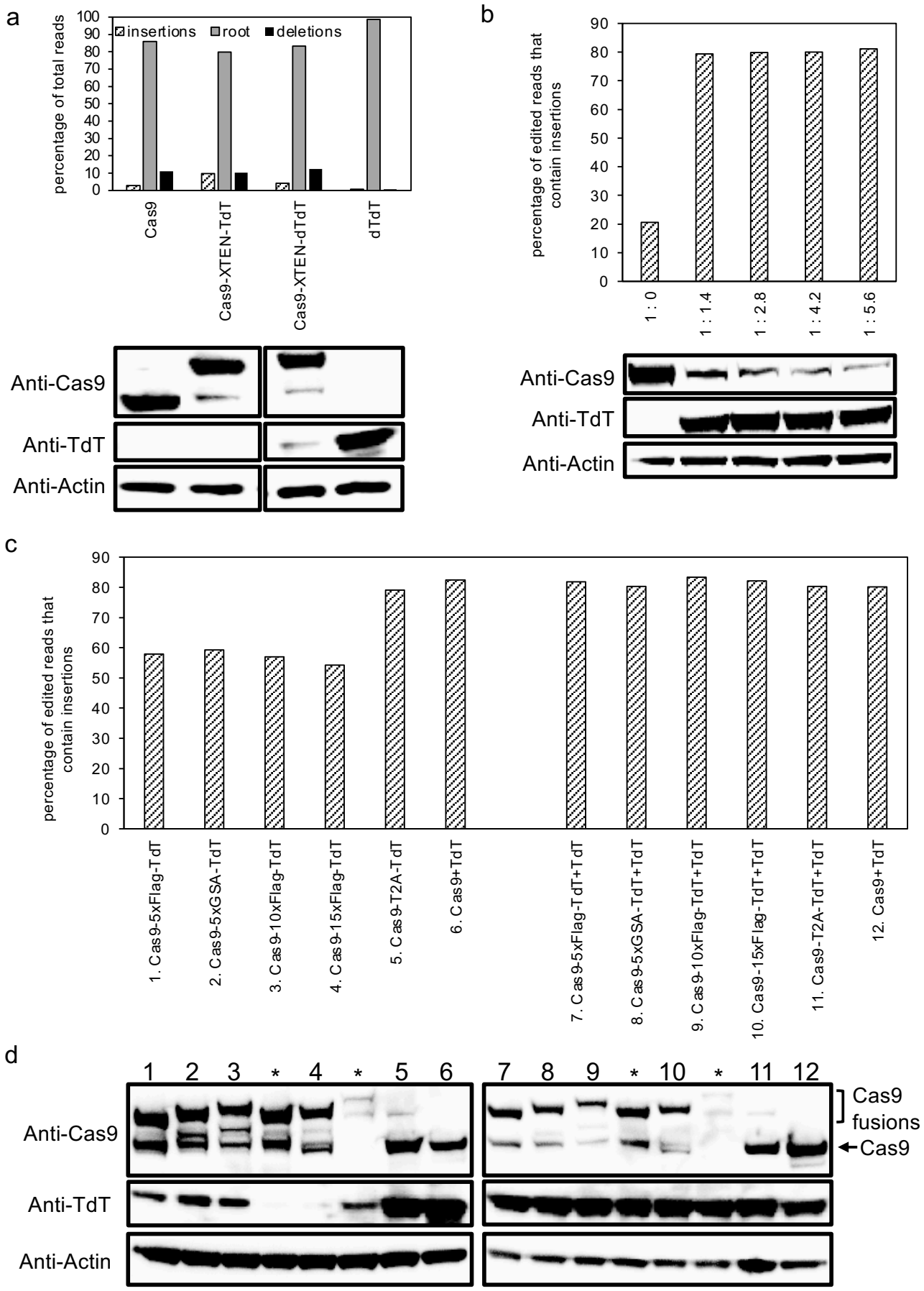




Insertion sequences for the 50%, 65%, and 80% GC-rich templates for the PCR bias assay of Q5 and Phusion polymerases.

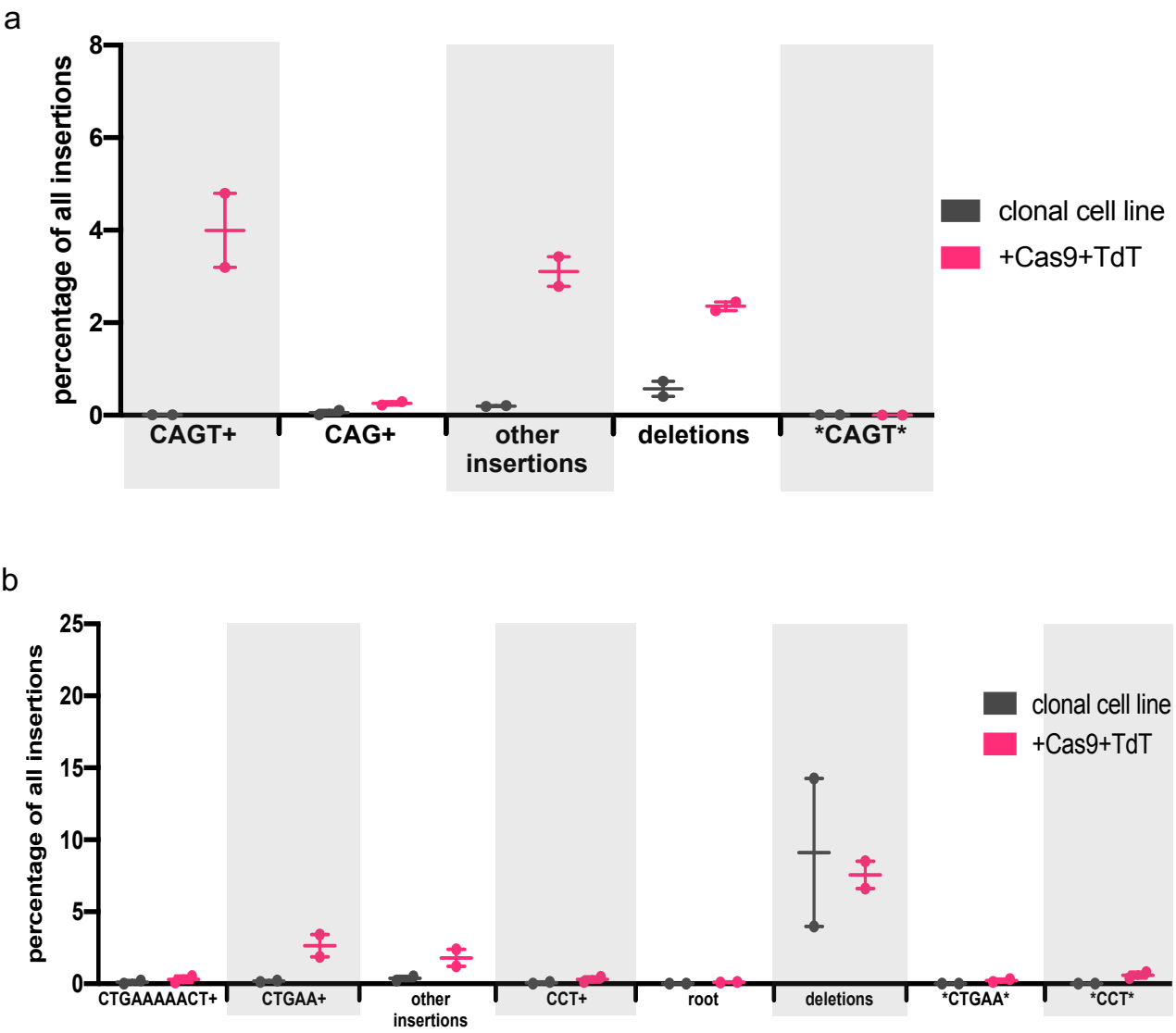
% GC	5'-Insertion Sequence-3'
50	CATGGCATGGATCCTGTGACTCGATATAGGAACTCCGATC
65	CCGTGTCAGCTCACTGCGATCCAGGGTCCACTCCCAGTGC
80	CCGTGGCAGCTCGCTGCGCGCCAGGGTCCACGCCCAGGGC

Supplementary Figure 1. Expression of Cas9 upon transfection, effect of TdT on editing outcomes at a variety of genomic sites, and test of fidelity of NGS library prep for GC-rich templates. Related to Figure 2 and Methods. (a) Western blot of samples from Figure 2a. **(b)** Percentages of deletions and insertions at different -4/-3 nucleotide cut sites. HEK293T cells were transfected with the appropriate sgRNA and either a Cas9 or Cas9-T2A-TdT construct. Bars represent the average of two biological replicates. **(c)** Percentages of the total number of amplicons detected by NGS from a 1:1:1 molar ratio of the 50%, 65%, and 80% GC-rich templates. Bars represent the average of two technical replicates.

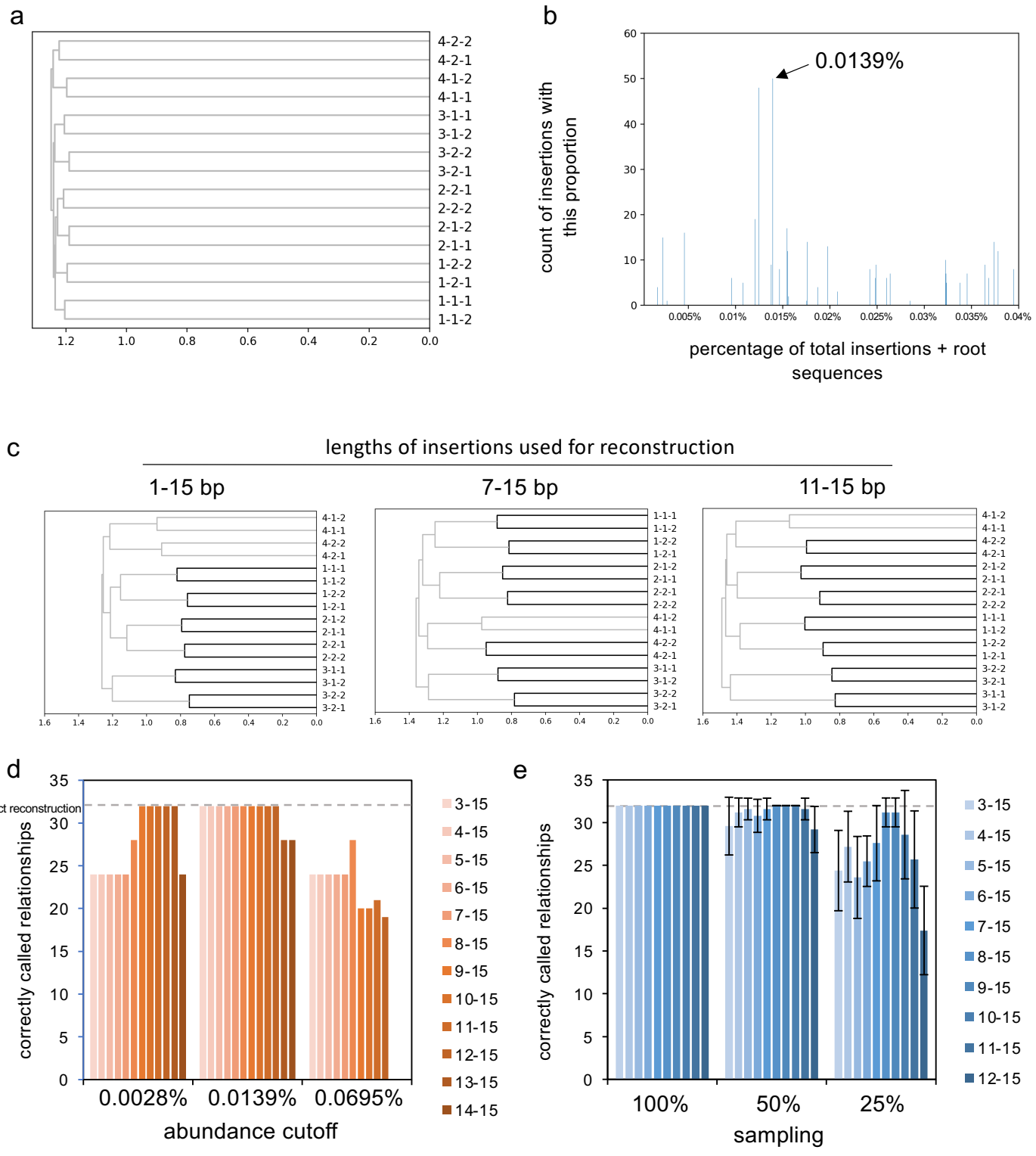


Supplementary Figure 2.

Supplementary Figure 2. Activity of varying amounts and fusions of TdT. Related to Figure 2. (a) TdT promoted increased insertions at Cas9 cut sites *via* its polymerase activity. A catalytically dead version of TdT, "dTdT," was created by introducing the mutations D343E and D345E. 293T cells were transfected with plasmids expressing Cas9, Cas9-XTEN-TdT, Cas9-XTEN-dTdT, or dTdT alone, and an sgRNA against a genomic site (HEK293site3). Three days later, cells were collected, DNA was extracted, and the targeted genomic site was amplified by PCR and sequenced by NGS. Sequences were annotated as unchanged (root); pure insertions (insertions); or any sequence that leads to a loss of information (deletions). Western blots were performed on equal amounts of protein from each sample. This experiment was performed once. **(b)** An increased ratio of TdT to Cas9 near our current range did not increase insertion rate. Plasmids expressing Cas9 and TdT were mixed in the ratios noted and transfected into 293T cells along with an sgRNA against a genomic site (HEK293site3). Three days later, cells were collected, DNA was extracted, and the targeted genomic site was amplified by PCR and sequenced by NGS. Sequences were annotated as unchanged or changed. Then, the proportion of changed sequences that were pure insertions was calculated. Bars represent the average of two technical replicates. Western blots were performed on equal amounts of protein from each sample. Actin serves as a loading control for Cas9 and a sample-processing control for TdT. **(c)** Expressing Cas9-TdT fusions promoted insertions, but less efficiently than free TdT. Plasmids expressing various Cas9-TdT fusions were transfected into 293T cells along with an sgRNA against a genomic site (HEK293site3), with or without additional free TdT. Three days later, cells were collected, DNA was extracted, and the targeted genomic site was amplified by PCR and sequenced by NGS. Sequences were annotated as unchanged or changed. Then, the proportion of changed sequences that were pure insertions was calculated. Bars represent the average of two technical replicates. **(d)** Western blots were performed on equal amounts of protein from the single biological replicate of each sample shown in (c). * represents fusion proteins not included in the analysis.

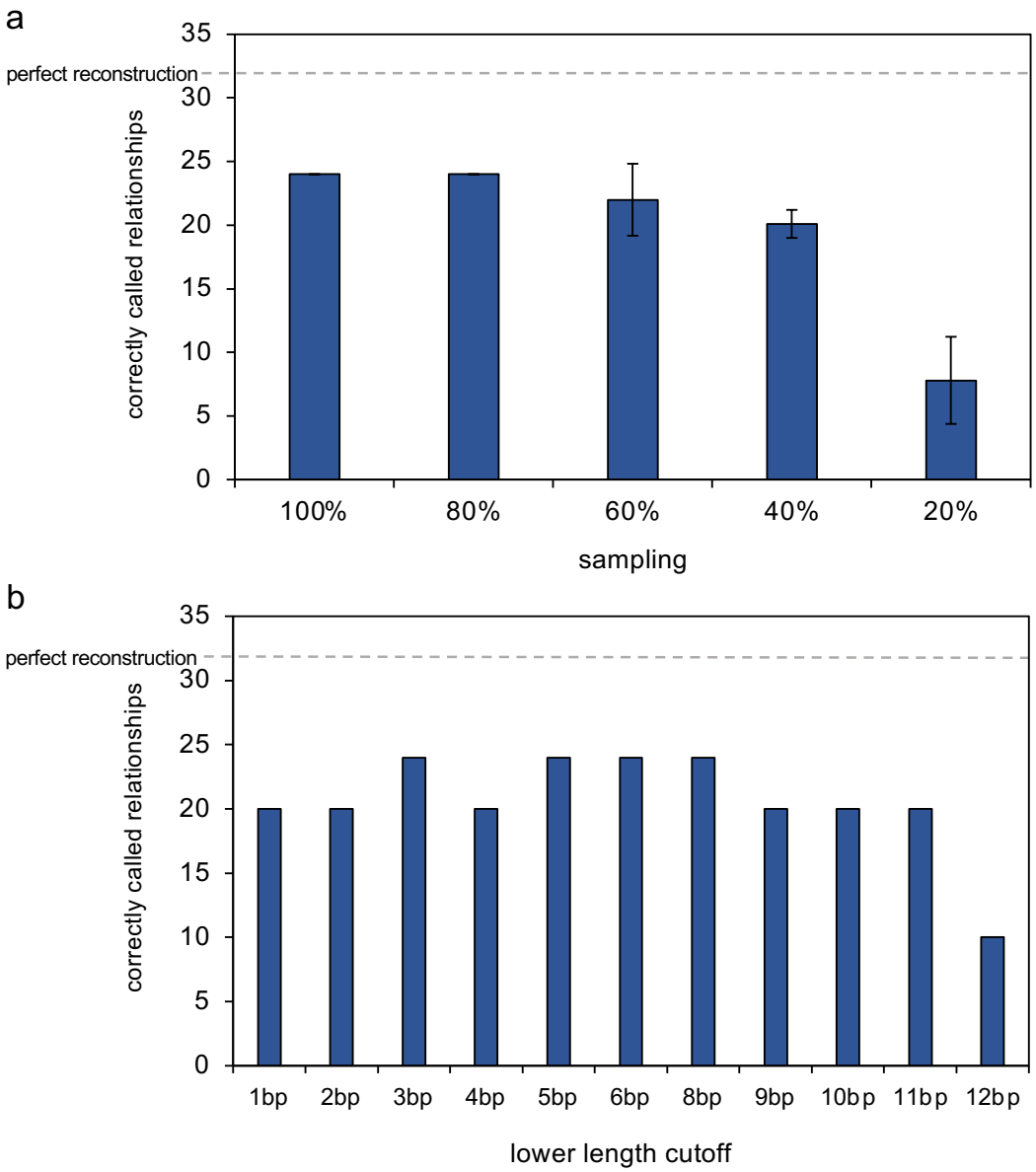


Supplementary Figure 3. Cas9 and TdT mediated multiple rounds of editing on an integrated hgRNA. Related to Figure 3d and Extended Data Figure 5. (**a**) 6-day timepoint for the experiment shown in Figure 3d. (**b**) 6-day timepoint for the experiment shown in Extended Data Figure 5b.



Supplementary Figure 4.

Supplementary Figure 4. High-information recording and optimal analysis is required when sampling is limited. Related to Figure 5. This Figure contains analyses of the lineage tracing data set shown in Figure 5. **(a**) Lineage relationships were correctly identified when all insertion sequences were used for reconstruction, without any count or length cutoffs. Reconstruction was performed by determining Jaccard similarities between pairs of wells and then performing UPGMA hierarchical clustering, as in Figure 5b. **(b**) The abundance cutoff for this experiment was determined using a smaller, deeply-sequenced dataset that was obtained using the same library prep protocol. The number of insertions with each abundance was plotted, and 0.0139% was chosen as the cutoff. **(c)** Lineage reconstruction changed little as length cutoffs were changed. Reconstruction was performed as in Figure 5b. Insertions of the indicated lengths were used for reconstruction. **(d)** Abundance cutoffs affected reconstruction. Lineage reconstructions were performed by determining Jaccard distances between pairs of wells and then performing UPGMA hierarchical clustering, using insertions within the abundance and length cutoffs indicated. Reconstructions were scored as in Extended Data Figure 9a. Because the relationships between 16 wells were reconstructed, a maximum of 32 points is possible. Bars are missing for the most-stringent combinations of length and abundance cutoffs because there were not enough insertions above the cutoffs to calculate the Jaccard similarity. **(e)** Length cutoffs became more important as sampling efficiency declined. 50% or 75% of insertions were computationally removed at random from the dataset, then lineage reconstruction and reconstruction scoring were performed as in (d). Computational removal of insertions at random and subsequent reconstruction was performed 10 times for each insertion length cutoff and sampling efficiency. The mean reconstruction score for these 10 replicates is shown (error bars=$\pm$stdev).



Supplementary Figure 5. Further analysis of reconstruction of cell relatedness by DNA recording in primary cells. Related to Figure 5 and Extended Data Figure 8. (**a**) The data set was degraded by computationally removing unique insertions at random from each sample, so that only the indicated proportion of unique insertions remained. This computational degradation process was performed 10 times. The number of correctly reconstructed relationships was calculated as in Extended Data Figure 9a. Bars represent the mean of 10 replicates, using insertions of length 7-15 bp for the reconstruction (error bars=±stdev). (**b)** Reconstructions were scored for each minimum insertion length used for reconstruction.