bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Lineage tracing using a Cas9-deaminase barcoding system targeting endogenous L1 elements

Byungjin Hwang[1, †], Wookjae Lee[1, †], Soo-Young Yum[2, †], Yujin Jeon[1], Namjin Cho[1], Goo Jang[2, *], Duhee Bang[1, *]

[1]Department of Chemistry, Yonsei University, Seoul, Korea, [2]Laboratory of Theriogenology and Biotechnology, Department of Veterinary Clinical Sciences, College of Veterinary Medicine, Seoul National University, Seoul, Korea

† denotes co-first authors, * denotes corresponding authors

ABSTRACT

Determining cell lineage and function is critical to understanding human physiology and pathology. Although advances in lineage tracing methods have provided new insight into cell fate, defining cellular diversity at the mammalian level remains a challenge. Here, we developed a genome editing strategy using a cytidine deaminase fused with inactive Cas9 (dCas9) to specifically target endogenous interspersed repeat regions in mammalian cells. The resulting mutation patterns served as a genetic barcode, which was induced by targeted mutagenesis with single-guide RNA (sgRNA), leveraging substitution events, and subsequent read out by a single primer pair. By analyzing interspersed mutation signatures, we show the accurate reconstruction of cell lineage using both bulk cell and single-cell data. We envision that our genetic barcode system will enable fine-resolution mapping of organismal development in healthy and diseased mammalian states.

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

Understanding the history of a cell is attractive to developmental biologists and genetic technologists because the lineage relationship illuminates the mechanisms underlying both normal development and certain disease pathologies. Researchers have developed a vast arsenal of robust genomic tools to interrogate cells. Traditionally, determining the history of individual cells has been accomplished using fluorescent proteins[1], Cre-*lox*P recombinase[2], somatic transposon events[3], and accumulated microsatellite mutations[4]. More recently, various cellular barcoding strategies have been developed[5–9]. To this end, the clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 genome editing system[10–13] has been primarily used to develop many cellular barcoding methods. However, creating repeat copies of array elements is difficult using methods that incorporate the CRISPR/Cas9 system, resulting in the need to introduce multiple exogenous DNA barcoding array elements into the genome to create a stable transgenic line for model organisms.

The majority of the human genome consists of repetitive elements that have been utilized for live imaging[14] and fetal aneuploidy testing[15]. Recently, researchers have shown that the editing of multiple endogenous retrovirus genes can be achieved without altering normal development[16]. Inspired by this evidence, we reasoned that multiple target regions in various interspersed sites in repeat elements could serve as barcodes for lineage analysis. We hypothesized that the induction of double-strand breaks (DSBs) in numerous endogenous interspersed sites present in the genome would have a toxic effect on cells[17] and affect cell lineage. Thus, we used the recently proposed base editing method involving dCas9 fused with cytidine deaminase (referred to as BE3 in previous literature[18] but hereafter referred to as targeted deaminase), which converts C:G base pairs to T:A base pairs in the 4–8 nucleotide region from the protospacer adjacent motif (PAM) on the distal side of the protospacer sequence without inducing DSBs.

Consequently, we developed a new cellular barcoding method for lineage tracing using dCas9 fused with cytidine deaminase to target the long interspersed nuclear element-1 (L1) in the genome. We believed that the interspersed target regions of the targeted deaminase system could be leveraged

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to introduce various substitution patterns into the regions using only a single-guide RNA (sgRNA). We confirmed that these unique marker combinations were gradually introduced and that tracking the patterns enabled the accurate reconstruction of the lineage relationship. In addition, we validated our approach at the single-cell level using time-lapse imaging experiments as ground truth data. We expect that our genetic barcoding system could provide insight into normal cell development as well as molecular pathology.

RESULTS

Characterization and selection of target regions for the cell barcoding system

First, we evaluated the amplified regions that were known to be human L1 retrotransposon regions[15]. A single primer pair was used to amplify the L1 retrotransposon region (**Figure 1a and Supplementary Figure 1a**). The region was designed to maximize the number of distinct sequences and to increase the potential for uniform amplification[15]. As expected, the amplicon size was bimodally distributed, with 99% of the regions overlapping with known L1 subfamilies (**Supplementary Figure 2,** see Methods).

Next, we searched for the endogenous regions that could serve as barcodes after genome editing using targeted deaminase. Like the existing CRISPR/Cas9 system, targeted deaminase targets a 20 bp spacer adjacent to the 'NGG' PAM sequence and induces C-to-T conversion in the specific protospacer sequence window. For proof-of-concept, we screened all targetable spacer regions to identify those with the 'NGG' PAM sequence in the retrotransposon region and compiled the candidate spacer regions with a C in the 4–8 nucleotide window from PAM on the distal side of the protospacer. We confirmed that there were several sgRNA candidates with identical spacer sequences that satisfied our established conditions. We then selected two spacer sequences with the highest number of precisely matched sites (top two, highest ranking [sgRNA-1] and third ranked [sgRNA-3]) in the targeted repeat regions (**Supplementary Figures 1b and 3**). The sequences of sgRNA-2 and

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sgRNA-1 were nearly identical (only a single base difference); thus, sgRNA-2 was omitted from further experiments. We expected that one sgRNA would introduce a substitution at multiple target sites in the L1 regions and defined 'cell barcode' as C-to-T substitutions (see Methods). Although the multiple target sites of each sgRNA had identical spacer sequences, they were distinguishable after amplification of the target region via the surrounding sequences, which could be uniquely aligned to the specific genomic position (**Figure 1b**).

Analysis of genome editing in repeat regions

We first applied the targeted deaminase system (**Supplementary Table 1 and Supplementary Figure 4**) on L1 retrotransposon regions in HEK293T and HeLa cells to test whether the system could be used in lineage tracing experiments. When we analyzed multiple target sites (targeted by a sgRNA-1) with identical spacer regions, the average editing efficiencies (fraction of the number of C>T reads) for HEK293T and HeLa cells were 1.5% and 2.3%, respectively (114 and 143 cell barcodes, respectively), for multiple targets in the known 4–8 nucleotide spacer sequence windows (**Figure 1c**), implying a low number of edited cells. In contrast, sgRNA-3 exhibited an average 4-fold higher editing efficiency (6.3% and 9.3% in HEK293T and HeLa cells, respectively) compared with sgRNA-1, and the correlation of the editing efficiency between the two 'C's in the spacer sequence windows were high in the multiple targets (**Figure 1d**). We also observed differences in the editing efficiency between the two cell lines across the target sites. No significant background mutations were detected compared with the vehicle control without targeted deaminase and sgRNA (*P*-value < 2.2e-16, Mann-Whitney U test). As well, aside from Cs in the known 4-8 nucleotide window in the PAM distal end, very low editing frequency of non-target Cs was observed (*P*-value < 2.2e-16, Mann-Whitney U test, **Supplementary Figure 5**).

Cumulative introduction of cell barcodes using targeted deaminase

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Next, we investigated whether the targeted deaminase system could continuously introduce genetic barcodes in the target repeat regions. To explore the extent of the continuous editing strategy at known time points, we compared sgRNA-1 and sgRNA-3 via repeat serial transfection of the targeted deaminase system at approximately 3-day intervals (**Figure 2a)**. The average editing efficiency increased linearly with the gradual accumulation of edited sites following serial transfection, and the observed editing rate was faster using sgRNA-3 compared with sgRNA-1. Thus, we concluded that the continuous introduction of genetic cell barcodes was feasible using our method.

Lineage tracing at the bulk level using a controlled *in vitro* cell line experiment

We devised an *in vitro* cell expansion experiment to investigate whether the cell barcodes were properly introduced each generation for the tree construction. HEK293T and HeLa cells were transfected using the PiggyBac™ transposon system containing the mCherry fluorescence protein, targeted deaminase, and sgRNA (**Supplementary Figure 4**). After sorting by FACS into 96-well plates, single-cells from the two cell lines underwent clonal expansion. The mCherry-positive single-cells were then sorted and cultured in different wells, and the process was repeated (**Figure 2b**). The target regions from the expanded cells were amplified using a single primer pair and subjected to next-generation sequencing (NGS). The system was first applied to the HEK293T cells. The known tree topology within each node (a node represents the sum of the cell barcodes) represented by the bulk cells enabled the validation of the lineage tracing. On average, 95% of the unique reads were aligned and these reads were processed further as multiple alignments can occur because of homology between the flanking regions. After alignment and variant calling, we detected an average of five cell barcodes per node in the first generation of the tree and found that ~93% of the target sites were unedited by sgRNA-1.

Next, tree building was performed using an iterative graph approach[7] and an additional *post-hoc* cell barcode selection step (see Methods). For sgRNA-1, the tree could not be identified correctly

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

because of a low number of informative cell barcodes. Conversely, sgRNA-3 showed an average of 29 cell barcodes for the first generation nodes. The accuracy of the reconstructed tree was defined based on the fraction of the correct node placements on the known depth and position of the tree. For sgRNA-3, we achieved 81% accuracy (29/36) of the reconstructed tree based on the variant calling approach. To remove spurious connections and refine the reconstructed tree, we developed a custom algorithm to obtain confident cell barcodes. We used a probabilistic approach with a posterior calculation to select the final candidate cell barcode for the tree construction (see Methods). Compared to the conventional variant calling approach, we observed an average of 70 cell barcodes for sgRNA-3 and the reconstruction accuracy improved to 92% (33/36) (**Figure 2b**). For the HeLa cells, we also observed slightly improved performance in accuracy (88% and 93% for conventional variant calling vs. custom algorithm, respectively) (**Supplementary Figure 6**).

Reconstruction of the lineage relationship at the single-cell level

Next, we explored whether our targeted deaminase system could be extended to the reconstruction of lineage relationships at the single-cell level. Because the tree reconstruction efficiency was better with sgRNA-3 than sgRNA-1, we focused on sgRNA-3 only and chose HeLa cells for the single-cell level lineage tracing for the ease of isolating single-cells. HeLa cells were transfected with the PiggyBac™ transposon system containing the mCherry fluorescence protein, targeted deaminase, and sgRNA-3. We used time-lapse imaging to generate ground truth tree data and picked individual cells (n = 32 total single-cells analyzed [3 different trees]) determined to be mCherry marker-positive by manual inspection (**Figure 2c**). To prepare for the sequencing experiment, we first performed whole-genome amplification (WGA) and subsequent PCR amplification of the selected single-cells. However, uneven distribution of the sequencing reads prohibited robust cell barcode identification. Furthermore, it has been reported that increased background C-to-T mutations can occur because of the high denaturation temperature during WGA[19]. Thus, we elected to optimize the single-cell PCR conditions. After the optimization, we achieved

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

more uniform distribution of depth over the target regions (**Supplementary Figure 7** and Methods). A standard agglomerative hierarchical clustering approach was used for three different experimental trees for three to four divisions (8-16 cells) of the expansion. Confident cell barcodes were encoded in the binary state and pairwise cell-to-cell distances were calculated for the tree reconstruction. Hierarchical clustering using an average-linkage method consistently recovered the known trees validated in the imaging experiment (**Figure 2d and Supplementary Figure 8**) with a high computed Cophenetic correlation (0.92, 0.91, and 0.81).

To determine if the editing rate could affect the accuracy of the lineage reconstruction generated using our platform, we first estimated the accumulated mutation rate by approximating the editing dynamics of sgRNA-3 (**Supplementary Figure 9**). After exponential fitting, a rate of 0.06 edits per hour accurately reflected the experimental editing dynamics. Using this parameter as the rate of editing efficiency, we conducted a simulation to predict the number of cell barcodes per cell expansion using an empirical number of target sites based on the experimental results. The simulation results indicated that the number of barcodes generated at each depth of the tree accurately reflected the experimental results obtained from the single-cells (comparison of the average number of cell barcodes per cell; experimental vs simulation; Tree 1: 6.4 vs 6.1, Tree 2: 20.3 vs 18, Tree 3: 9 vs 9.5, **Figure 2d and Supplementary Figure 8**). Using a model that included the number of target sites and editing rate as variables, we found that the accuracy (Cophenetic correlation) improved with an increasing number of target sites and editing rate (**Supplementary Figure 10**).

DISCUSSION

Here, we demonstrated a lineage analysis platform using targeted deaminase with sgRNA. Further, we showed that library preparation could be simplified using a single flanking primer pair to amplify the edited endogenous repeat element (L1 retrotransposon region[20]). As well, our method relies on more predictable barcodes based on substitution patterns. Most importantly, our method does

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

not rely on introducing exogenous barcodes without making dsDNA breaks. We used endogenous repeat elements as potential barcodes, eliminating the need to create and insert complex repeat arrays. The genetic editing pattern in endogenous sites can serve as a cell barcode to reconstruct the cell lineage tree. A limitation of conventional Cas9-induced lineage tracing is the rapid saturation of genetic barcodes, which might hinder long-term labelling experiments. In contrast, the low editing rate of our method (0.06 vs 0.4 per hour[7] for our method vs conventional Cas9-induced method, respectively) enables long-term lineage tracing and can be further controlled using different sgRNA designs as demonstrated by the observed differences between sgRNA-1 and sgRNA-3. Notably, the high theoretical diversity (sgRNA-3, 10^{250}) of the cell barcodes is sufficient to cover approximately 37 trillion cells in the human body. Although we have only used targeted deaminase for the lineage analysis, the mutation rate could be controlled using other base editing systems[21,22] or the rational design of sgRNA that target other regions. In addition, other engineered dCas9 proteins or other classes of CRISPR proteins[23–26] might influence the editing efficiency or target diversity. Alternatively, we could simultaneously integrate multiple sgRNAs into the genome and track the resulting mutation patterns to reconstruct the lineage in a more predictable manner.

Our method has some technical challenges. Although the potential diversity of our proposed method is substantially larger than the diversity achieved using previous methods, we could not fully explore this aspect because of the limited number of single-cells sampled and the low editing rate. In the conventional Cas9 system, edited barcode arrays are amplified and can be readout by paired-end sequencing[5]. Conversely, clonal information could not be directly inferred from our bulk cell analysis because of the interspersed nature of the barcode regions. The analyzable clonal sequence space is likewise limited by current sequencing length (300 bp for the Illumina platform using the 150 bp paired-end mode). Instead, we devised a lineage relationship between the bulk cells (isolated from a single well) that could be determined based on the integrated cell barcode pattern of the cell population.

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We employed a custom algorithm to select robust cell barcodes for bulk cell-based lineage tracing because conventional variant calling based on the stringent FreeBayes algorithm returned a small number of variants. This allowed flexible control of the various parameters and eliminated the need for simple, sequencing depth-based filtering. In a previous study, the original tree topology was preserved with high accuracy using a graph-based approach[7]. In our single-cell analysis, tree building for three different sets of three to four generation single-cell division was also accurate. In simulations, we detected higher accuracy of the tree reconstruction with an increasing number of barcodes. In a practical scenario in which conventional Cas9 causes barcode dropout, our base editing scheme (without dsDNA breaks) could potentially be useful for lineage construction. In the future, additional comparative studies are required to investigate whether the base editing strategy confers an advantage over the Cas9 nuclease-based method from the perspective of the effect of barcode dropout on tree reconstruction.

In summary, we have shown a proof-of-concept for using the targeted deaminase system for lineage tracing through *in vitro* tree expansion experiments and time-lapse imaging. In doing so, we have laid the foundation for developing an alternative technique to trace cell lineages using endogenous targets. In future studies, we hope that our system can be used to determine the fates of different cells for organ development in a transgenic model organism. In the long-term, if coupled with transcriptional information, our method could provide a high-resolution view of developmental lineages.

Methods

Plasmid construction

The PB CMV-BE3 EF1α-mCherry-T2A-puro sgRNA (**Supplementary Table 1**) was constructed by PCR assembly of the U6-sgRNA expression cassette from the gRNA_Cloning Vector (Addgene plasmid #41824, Addgene, USA), whereas CMV-BE3 was constructed from pCMV-BE3

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Addgene plasmid #73021), the puromycin gene from lentiGuide-Puro (Addgene plasmid #52963), and mCherry from PB_tet_attB-mCherry (a gift from Seung Hyeok Seok in Seoul National University College of Medicine) on a PiggyBac transposon backbone from PB-CA (Addgene plasmid #20960). The targeting sgRNA sequences were cloned according to the provided hCRISPR gRNA synthesis protocol. The pCy43 PiggyBac Transposase vector was provided by the Sanger Institute (Hinxton, UK). All cloned plasmids were confirmed by Sanger sequencing. Plasmids were prepared using the EndoFree Plasmid Maxi Kit (QIAGEN, USA) and Exprep Plasmid SV kit (GeneAll, Korea) according to manufacturers' protocols.

Cell lines and cell culture

All cell lines were obtained from the KCLB (Korean Cell Line Bank) and maintained at 37°C with 5% CO_2. The HEK293T human embryonic kidney and the HeLa human cervical cancer cell lines were cultured in Dulbecco's Modified Eagle's Medium (DMEM; Gibco, USA) supplemented with 10% fetal bovine serum (FBS; Gibco, USA) and 1% penicillin/streptomycin (P/S; Thermo Fisher Scientific, USA).

Generation of cells containing the genetic barcoding system

To generate cells harboring genetic barcodes, HEK293T and HeLa cells were individually transfected with genetic barcoding system plasmids at a 2:1 transposon (PB CMV-BE3 EF1α-mCherry-T2A-puro sgRNA) to transposase ratio using Lipofectamine™ 3000 (Life Technologies, USA) according to the manufacturers' protocols. The transfected cells were incubated for approximately 3 days and harvested for genomic DNA (gDNA) using the DNeasy Blood & Tissue Kit (QIAGEN, USA).

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The experiment that entailed the continuous introduction of genetic barcodes was conducted using the aforementioned transfection method. Half of the cells were harvested for gDNA 3 days after transfection and the other half of the cells were cultured for the serial transfection.

In the experiment to measure barcode editing efficiency, only the genetic barcoding system without transposase was transfected using Lipofectamine™ 3000. The transfected cells were harvested at 4, 8, 12, 16, 20, 24, 30, 36, 42, and 48 hours, followed by gDNA extraction.

A controlled *in vitro* tree experiment

The procedures used to generate the barcodes were identical to the protocols described above, and successfully transfected cells were selected using puromycin (2 μg/ml). The selected cells were isolated by single-cell FACS using an Aria II FACS machine based on mCherry fluorescence (a marker of transfection). The sorted single-cells were then cultured in DMEM supplemented with 20% FBS and 1% P/S. Single-cell-derived populations were expanded in culture for up to 3 weeks. Half of the expanded cells were harvested and gDNA was extracted for NGS as described in the 'Endogenous genetic barcode amplification – bulk cells' section and the remainder of the cells from each lineage were pelleted and frozen at -20°C.

Endogenous genetic barcode amplification – bulk cells

The gDNA extracted from the cells described above were used for the amplification of endogenous genetic barcodes. Kapa High Fidelity Polymerase (Kapa BioSystems, USA) was used for all barcode amplifications. Up to 500 ng of gDNA was loaded into a single 20 μl initial PCR reaction that included 1 μl of 10 μM forward and reverse primers (L1 site for and L1 site rev in **Supplementary Table 2**), 10 μl KAPA DNA polymerase, and up to 8 μl of nuclease-free water and amplified using primers with a sequencing adaptor according to following protocol: 98°C for 120 s

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

followed by two cycles of 98°C for 10 s, 57°C for 120 s, and 72°C for 120 s and a final 10 min step at 72°C. Homemade AMPure XP beads[27] (hereafter, AMPure beads) generated using Sera-Mag SpeedBeads (6515-2105-050350, Thermo Scientific, USA) were then used to purify the initial PCR products. The initial PCR products were then loaded into a single 20 μl second index PCR reaction and amplified with index primers using the following protocol: 30 s at 98°C followed by 15 cycles of 10 s at 98°C, 30 s at 60°C, and 30 s at 72°C and a final10 min step at 72°C. The second PCR product was then purified using 1.2× AMPure beads. All primers were prepared by IDT (Integrated DNA Technologies, USA). The sequencing was performed on the Illumina NextSeq 500 system using a NextSeq 500/550 High Output v2 kit (300 cycles) (Illumina, USA). The sequencing statistics for the bulk experiments are provided in **Supplementary Table 3**.

Transfection for time-lapse imaging

The HeLa cells were prepared by subculturing in a 35-mm dish with 2 mL DMEM, supplemented with 1% P/S, 1% non-essential amino acids (NEAA) (Gibco, USA), 100 mM 2-mercaptoethanol (2-ME) (Sigma-Aldrich), and 10% FBS. The cells were transfected with plasmids using an electroporation system (Neon®, Invitrogen, voltage:1140 v, pulse width range: 40 ms, pulse number: 1). After 4 hours of transfection, the culture medium was replaced with fresh medium to remove the dead cells.

Manual cell picking and monitoring for time-lapse imaging

All of the single-cell manipulations were conducted using a micromanipulator device (Nikon-Narishige, Tokyo, Japan) during observation under an inverted microscope. One day after transfection, the cells were trypsinized and washed in phosphate-buffered saline (PBS; Gibco, USA). To manually pick single-cells, the cell suspensions were placed in a drop of PBS with 0.5% FBS and covered by mineral oil (Sigma, USA). A single RFP-positive only cell was aspirated using a micro-injection

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pipette (diameter: 20 μm, ORIGIO, Charlottesville, VA) under fluorescence exposure. The aspirated single-cell was transferred to a 4 μl droplet of DMEM supplemented with 1% P/S, 1% NEAA, 100 mM 2-ME, and 10% FBS in a 100-mm dish and overlaid with mineral oil at a ratio of one cell to each droplet. The single-cells were cultured for 4 hours in a CO_2 incubator at 37□ and then moved to an incubator equipped with a JuLI™ Stage real-time cell history recorder (NanoEnTek), which enables the observation of cell growth using live, time-lapse imaging.

Endogenous genetic barcode amplification – single-cell

Whole-genome multiple displacement amplification (MDA) of the single-cells (from manual cell picking) was conducted using the illustra™ Ready-To-Go GenomiPhi V3 DNA Amplification Kit (GE Healthcare, USA). MDA was performed according to manufacturers' protocols with the slight modification of increasing the reaction time from 1 hour 30 min to 3 hours. Next, 5 μl of the single-cell MDA product was added to a 20 μl initial adaptor PCR reaction that included 1 μl of 10 μM forward and reverse primers (L1 site for and L1 site rev in **Supplementary Table 2**), 10 μl KAPA DNA polymerase, and up to 8 μl of nuclease-free water. The PCR reaction was performed using the following protocol: 2 min at 98°C followed by 10 cycles of 10 s at 98°C, 2 min at 57°C, and 2 min at 72°C and a final 2 min step at 72°C. After purification using AMPure beads, the initial PCR product was loaded into a second index PCR reaction and PCR was performed using the following protocol: 30 s at 98°C followed by 15 cycles of 10 s at 98°C, 30 s at 60°C, and 30 s at 72°C and a final 10 min step at 72°C. The final product was then purified using AMPure beads.

Single-cell PCR amplification of the manually picked single-cells was performed using the PCR reaction composition ratio described above and following protocols: 1. Adaptor (initial) PCR: 2 min at 98°C followed by 30 cycles of 10 s at 98°C, 2 min at 57°C, and 2 min at 72°C and a final 2 min step at 72°C. The product was then purified using AMPure beads. 2. Index (second) PCR: 30 s at 98°C followed by 15 cycles of 10 s at 98°C, 30 s at 60°C, and 30 s at 72°C and a final 10 min step at

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

72°C. The products of the second PCR were then loaded onto a 2% agarose gel and separated by gel electrophoresis. The band at the expected size was purified using the QIAquick Gel Extraction Kit (QIAGEN, USA) according to the manufacturers' protocols. Likewise, sequencing was performed on the Illumina NextSeq 500 using the NextSeq 500/550 High Output v2 kit (300 cycles).

Analysis of amplified post-alignment processing

Reads were aligned to hg19 using BWA (v0.7.12-r1039) and realignment around indels (RealignerTargetCreator, IndelRealigner) was performed using GATK (v3.3-0). Per-position base calling was accomplished using the SAMtools (v1.1) *mpilup* function and the pileup file was used for custom variant calling (details in the next section). The aligned regions were annotated using RepeatMasker (http://www.repeatmasker.org) and the sizes of the amplified regions were plotted to calculate the overlap fraction.

Identification of confident sites for lineage reconstruction

We first adopted a variant calling approach using FreeBayes (v1.1.0-3-g961e5f3) to extract confident markers (C>T substitutions) for the lineage reconstruction. The variant calling used FreeBayes (input from BAM after indel realignment) and filtered positions (depth > 10) considered candidate markers, and only included the markers with higher allele frequency than the value calculated for the background control using an empty vector. For the bulk and single-cell linage tracing experiments involving HeLa cells, variant calling was performed using modified parameters (--ploidy 3, --pooled-discrete). To handle both the bulk and single-cell data efficiently, we developed a custom algorithm for a variant calling strategy that was based on our targeted deaminase system. We adopted a probabilistic approach using a binomial mixture model with conditional probabilities, as described in a previous study[28]. An expectation-maximization algorithm was used to estimate the model parameters to account for the inherent deviation of allele frequencies in unstable genomes (e.g.,

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genomes with different ploidies). Every candidate position in the target region, depth >10×, variant allele count > 2, and posterior probabilities ≥ 0.95 was selected as a final marker. After performing a union operation for all the markers present in the bulk nodes, we selected confident markers using following criteria: 1) sites having a minimum editing efficiency of 0.1 and 2) removal of highly correlated markers (≥0.9). The sum of these markers for each cell represented the final 'cell barcode.'

Tree building for bulk and single-cell experiments

For the bulk cell experiment, tree building was performed using an approach similar to the previous method (https://bitbucket.org/Bastiaanspanjaard/linnaeus). We used the base editing pattern to build the substitution graph. For simplicity, the nodes were identified as CIGAR string-like sequences (i.e., 1E10E means the first and the 10-th C position in the perfect on-target region was edited). Inherently, the first ancestor node (bulk cells) had a maximal degree of connections with the other nodes. Iteratively, we removed this node and identified the ancestor node from the remaining connected components. This procedure was repeated until all the nodes were designated. Once all the pairwise cell networks were built, the cells were placed in the graph. We did not use the 'cell doublet' detection threshold because scRNA-seq was not used in this study.

For the single-cell-based lineage tracing, the information was restricted regardless of whether the site was edited. To identify confident markers, blacklist candidate regions (integration of the single-cell results exhibiting no mCherry signal or vehicle control single-cells) were also filtered out. Unlike the bulk cell lineage construction, the time-lapse-based single-cell experiment contained the cells from the last depth of the expansion. Thus, the lineage tracing was accomplished using a different logic. The distance between the cells was calculated using the Jaccard index and hierarchical clustering was performed using the *pvclus*t function in R. Approximately unbiased probability values (p-values) were calculated based on 1,000 iterations.

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Inferring editing efficiency dynamics and simulations

The cells were transfected with the targeted deaminase vector as described above. After 4, 8, 12, 16, 20, 24, 30, 36, 42, and 48 hours, the bulk cells were collected and amplified for sequencing (two replicates). Assuming a wild type fraction of 1 (100%) in t=0, we estimated the editing rate (λ) by fitting the exponential function. We designated the wild type percentage as the fraction of unedited sites (C>T candidate positions) in the perfect on-target regions.

Data analysis

All statistical analyses were performed using the R computing environment and data plotting was conducted with the *seaborn* and *ggplot2* packages in R. For Figure 1c and Figure 2a, two-tailed Man-Whitney U test was used to test for significance of difference between the control and sample means.

Data availability

The sequencing data supporting the findings of this study are available in the Sequence Read Archive with the identifier SRA SRP151792.

Acknowledgements

This work was supported by: (i) the Mid-career Researcher Program (2015R1A2A1A10055972) from the National Research Foundation of Korea, funded by the Ministry of Science, ICT & Planning, (ii) the Bio & Medical Technology Development Program of the National Research Foundation (NRF), funded by the Korean government (MSIT; NRF-

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2016M3A9B6948494), and (iii) the Bio & Medical Technology Development Program of the NRF, funded by the Korean government (MSIT; NRF-2018M3A9H3024850).

Author contributions

D.B, B.H, and W.L conceived the project. B.H analyzed the data with input from Y.J. W.L performed the cell culture experiments with the help of N.C. S.Y performed the time-lapse imaging experiment. B.H, W.L, and S.Y wrote the manuscript with input from all coauthors. D.B and G.J supervised the project.

Competing interests

The authors declare no conflicts of interest.

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Livet, J. *et al.* Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450,** 56–62 (2007).

2. Pei, W. *et al.* Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* (2017). doi:10.1038/nature23653

3. Evrony, G. D. *et al.* Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85,** 49–59 (2015).

4. Reizel, Y. *et al.* Colon stem cell and crypt dynamics exposed by cell lineage reconstruction. *PLoS Genet.* **7,** e1002192 (2011).

5. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science (80-.).* **353,** aaf7907 (2016).

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

7. Spanjaard, B. *et al.* Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* **36,** 469–473 (2018).

8. Schmidt, S. T., Zimmerman, S. M., Wang, J., Kim, S. K. & Quake, S. R. Quantitative Analysis of Synthetic Cell Lineage Tracing Using Nuclease Barcoding. *ACS Synth. Biol.* **6,** 936–942 (2017).

9. Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain by scGESTALT. doi:10.1101/205534

10. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* **32,** 347–355 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

11. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science (80-.).* **339,** 819–823 (2013).

12. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8,** 2281–2308 (2013).

13. Jinek, M. *et al.* A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science (80-.).* **337,** 816–821 (2012).

14. Chen, B. *et al.* Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell* **155,** 1479–1491 (2013).

15. Kinde, I., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Thomas, D. FAST-SeqS: A Simple and Efficient Method for the Detection of Aneuploidy by Massively Parallel Sequencing. *PLoS One* **7,** e41162 (2012).

16. Yang, L. *et al.* Genome-wide inactivation of porcine endogenous retroviruses (PERVs). *Science* **350,** 1101–4 (2015).

17. Riches, L. C., Lynch, A. M. & Gooderham, N. J. Early events in the mammalian response to DNA double-strand breaks. *Mutagenesis* **23,** 331–339 (2008).

18. 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).

19. Dong, X. *et al.* Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat. Methods* (2017). doi:10.1038/nmeth.4227

20. Bedrosian, T. A., Quayle, C., Novaresi, N. & Gage, F. H. Early life experience drives structural variation of neural genomes in mice. *Science* **359,** 1395–1399 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

21. Komor, A. C., Badran, A. H. & Liu, D. R. Editing the Genome Without Double-Stranded DNA Breaks. *ACS Chem. Biol.* acschembio.7b00710 (2017). doi:10.1021/acschembio.7b00710

22. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551,** 464–471 (2017).

23. Hu, J. H. *et al.* Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556,** 57–63 (2018).

24. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351,** 84–8 (2016).

25. Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163,** 759–771 (2015).

26. Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* (2015). doi:10.1038/nature14299

27. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22,** 939–946 (2012).

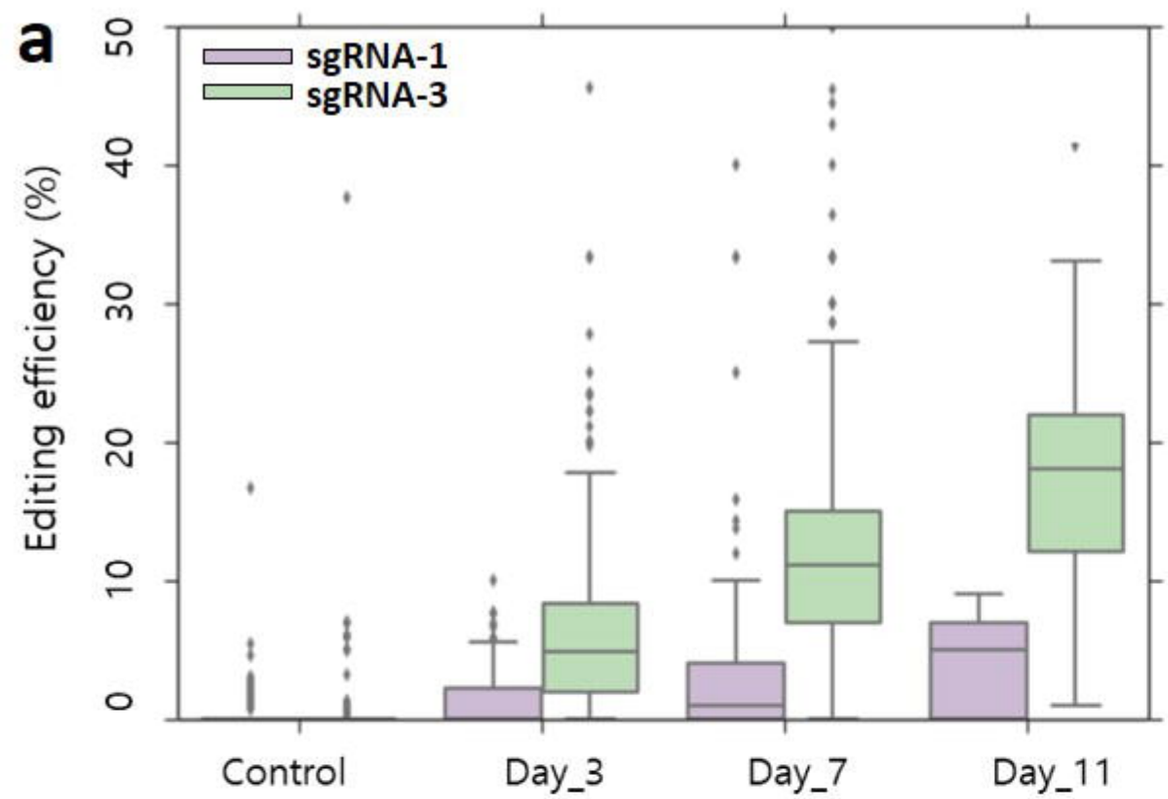
28. Goya, R. *et al.* SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26,** 730–6 (2010).

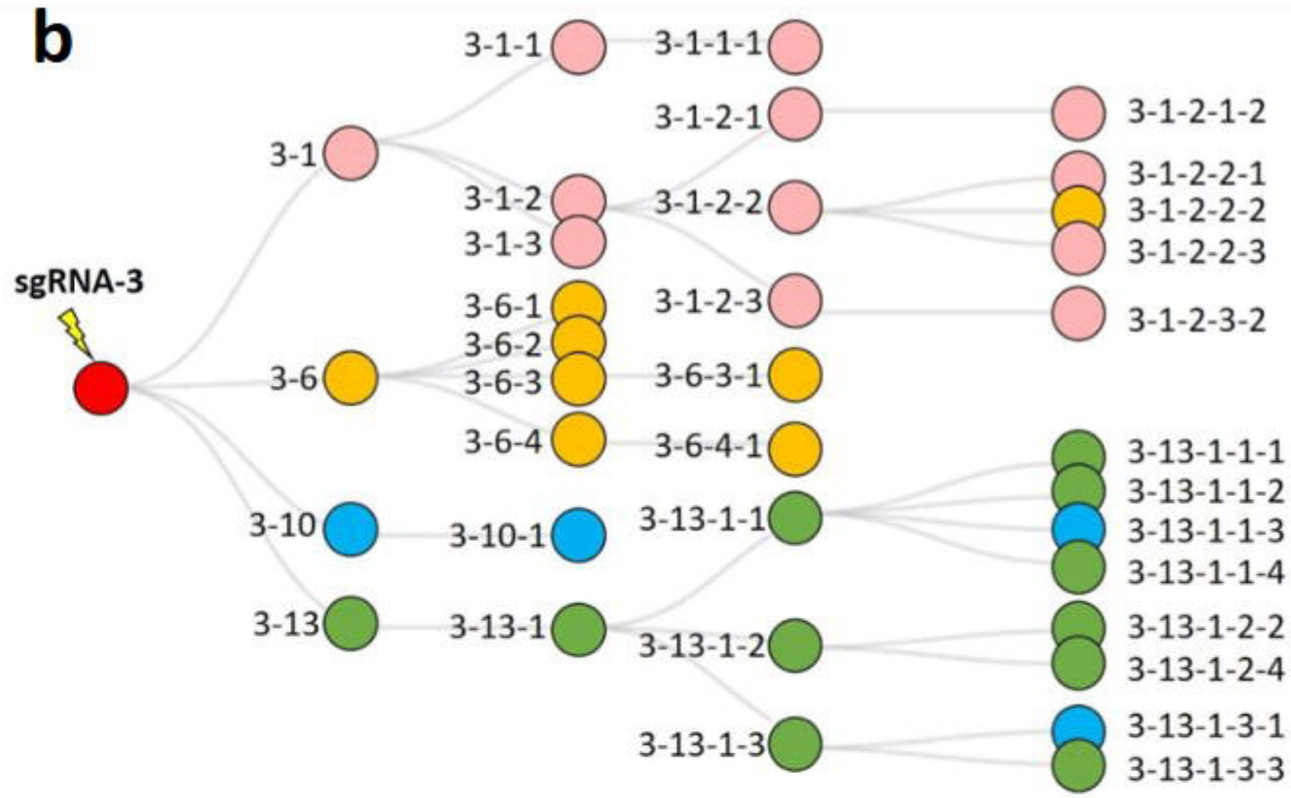
bioRxiv preprint doi: https://doi.org/10.1101/367177; this version posted July 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

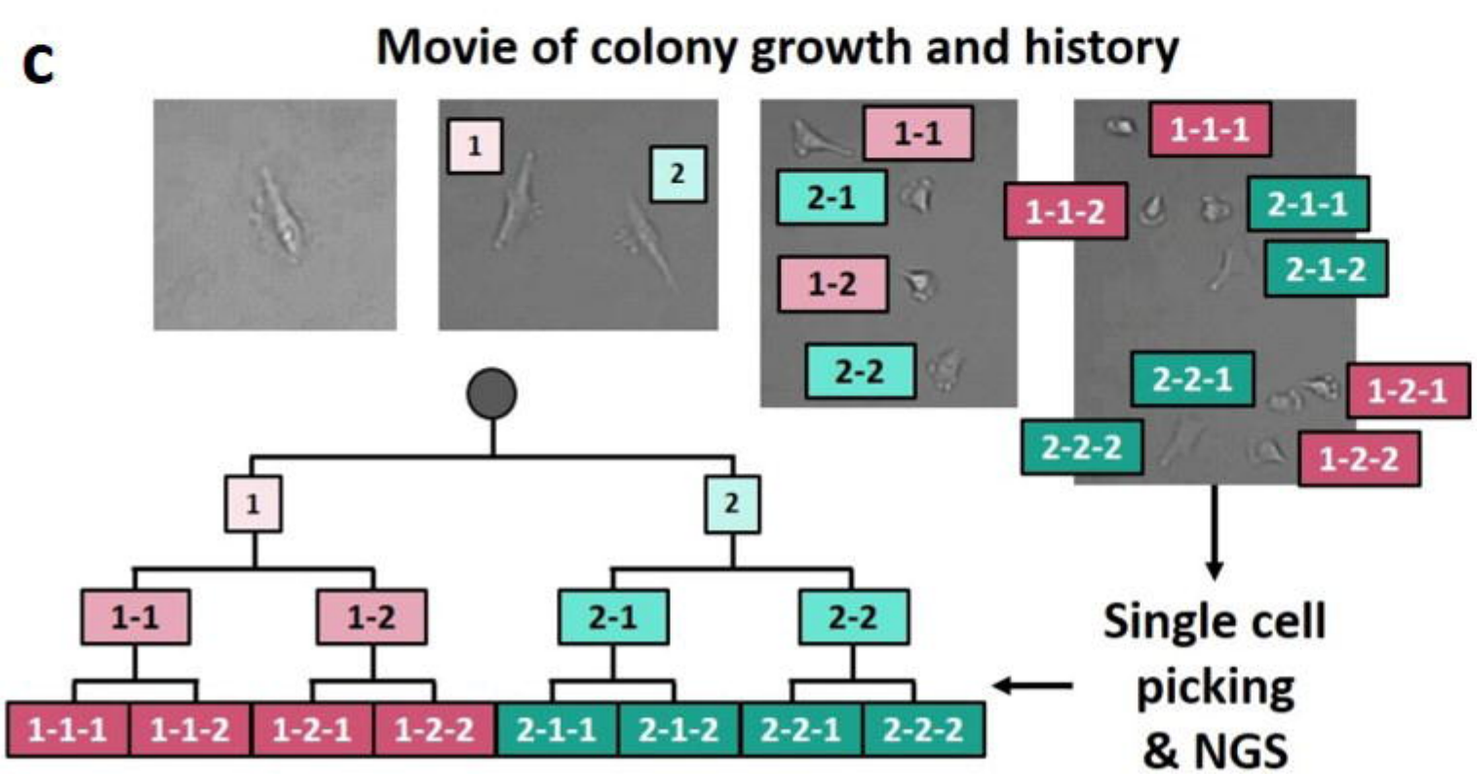
FIGURE AND LEGENDS

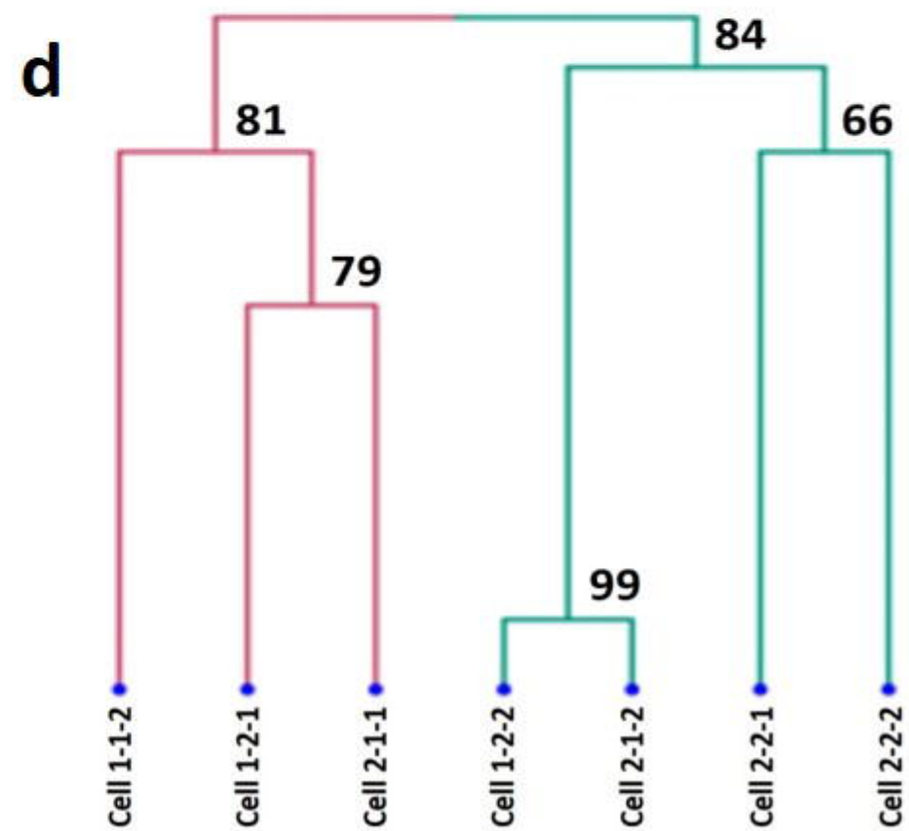
Figure 1. Targeted deaminase system for lineage tracing. a) Schematic overview of the targeted deaminase system. Red arrows indicate the single flanking primer pair used to amplify the targeted regions. The substitution pattern in the target region served as the cell barcode for tree reconstruction. b) Pairwise alignment of five representative target sites for sgRNA-1 and sgRNA-3 design. These regions were amplified using a single primer pair and alignment distinguished each target site by the different surrounding sequences. c) Editing efficiency of the two selected sgRNAs in the target region for HEK293T and HeLa cells. d) Correlation of editing efficiency between the first and second base 'C' in a specific window (4-8) within the sgRNA-3 spacer sequence in HEK293T cells (Spearman's correlation = 0.6 for HeLa cells).

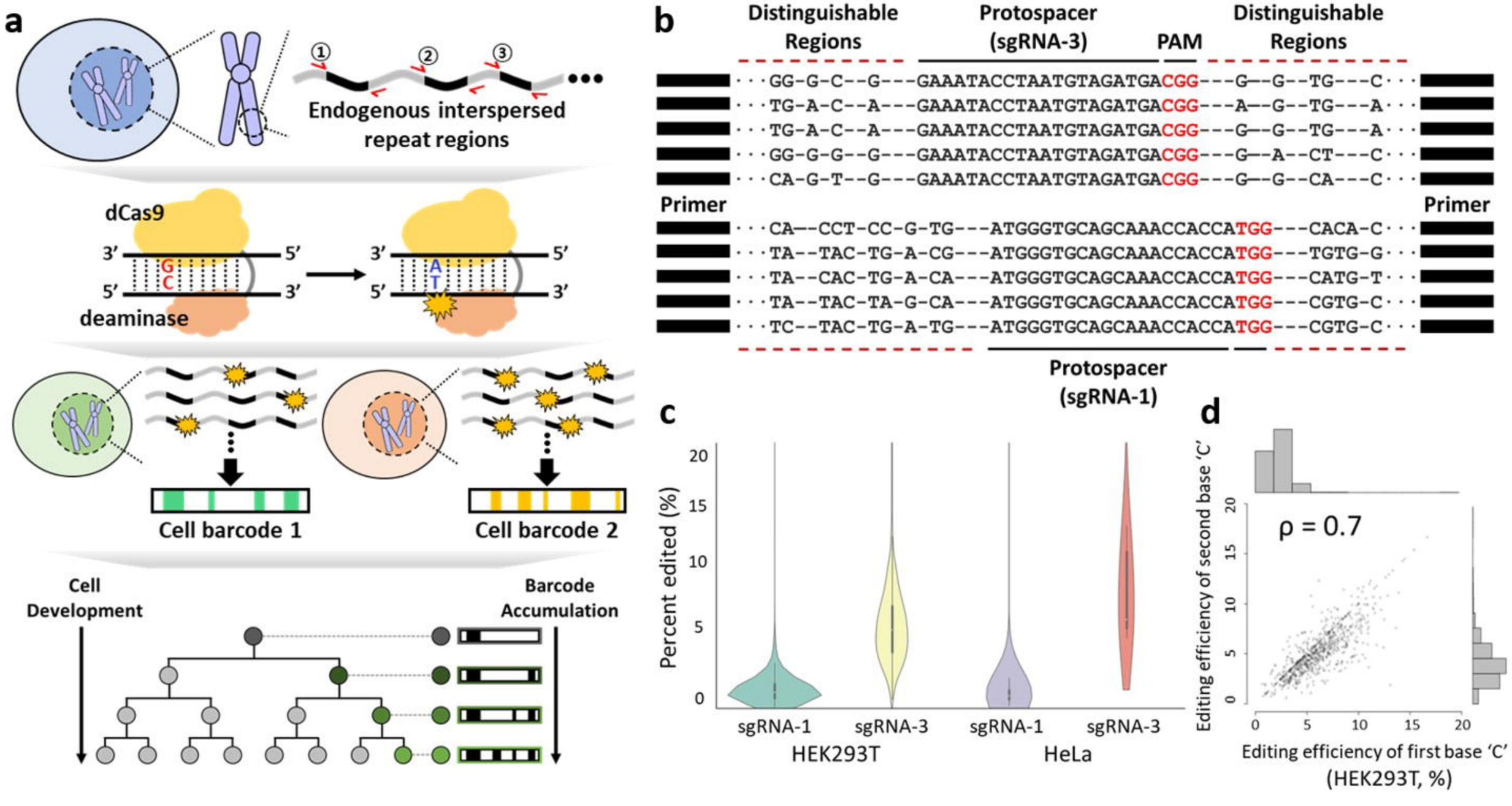
Figure 2. Cumulative editing and analysis of lineage tracing experiment. a) Cumulative editing efficiency in the target regions using sgRNA-1 and sgRNA-3 for the defined time points. b) A representation of the tree experiment for sgRNA-3 in HEK293T cells. The single-cells were sorted and expanded to a limited number of cell divisions. The bulk cells were then amplified from a 96-well plate and sequenced for cell lineage analysis. A network graph[7] was constructed from all the pairwise cell barcode information. The "-" represents the mother-child node relationship and the colors of each circle represent the different clade. c) Time-lapse imaging of the single-cell expansion experiment. At the end of the video, single-cells were picked and sequenced for lineage reconstruction. d) An example of a tree expansion experiment for seven single-cells. The bootstrapped confidence score (bootstrapped p-values × 100, n=1,000) are shown at the branching point. Physical isolation was impossible for the left most cell (1-1-1) in Figure 2c.


a
sgRNA-1
sgRNA-3
Editing efficiency (%)
0
10
20
30
40
50
Control
Day_3
Day_7
Day_11


b
sgRNA-3
3-1
3-6
3-10
3-13
3-1-1
3-1-2
3-1-3
3-6-1
3-6-2
3-6-3
3-6-4
3-10-1
3-13-1
3-1-1-1
3-1-2-1
3-1-2-2
3-1-2-3
3-6-3-1
3-6-4-1
3-13-1-1
3-13-1-2
3-13-1-3
3-1-2-1-2
3-1-2-2-1
3-1-2-2-2
3-1-2-2-3
3-1-2-3-2
3-13-1-1-1
3-13-1-1-2
3-13-1-1-3
3-13-1-1-4
3-13-1-2-2
3-13-1-2-4
3-13-1-3-1
3-13-1-3-3


c
Movie of colony growth and history
1
2
1-1
2-1
1-2
2-2
1-1-1
1-1-2
2-1-1
2-1-2
2-2-1
2-2-2
1-2-1
1-2-2
1-1-1
1-1-2
1-2-1
1-2-2
2-1-1
2-1-2
2-2-1
2-2-2
Single cell picking & NGS


d
84
81
66
79
99
Cell 1-1-2
Cell 1-2-1
Cell 2-1-1
Cell 1-2-2
Cell 2-1-2
Cell 2-2-1
Cell 2-2-2


a
Endogenous interspersed repeat regions
dCas9
deaminase
Cell barcode 1
Cell barcode 2
Cell Development
Barcode Accumulation
b
Distinguishable Regions
Protospacer (sgRNA-3)
PAM
Distinguishable Regions
Primer
Primer
Protospacer (sgRNA-1)
c
Percent edited (%)
sgRNA-1
sgRNA-3
sgRNA-1
sgRNA-3
HEK293T
HeLa
d
ρ = 0.7
Editing efficiency of second base 'C'
Editing efficiency of first base 'C' (HEK293T, %)