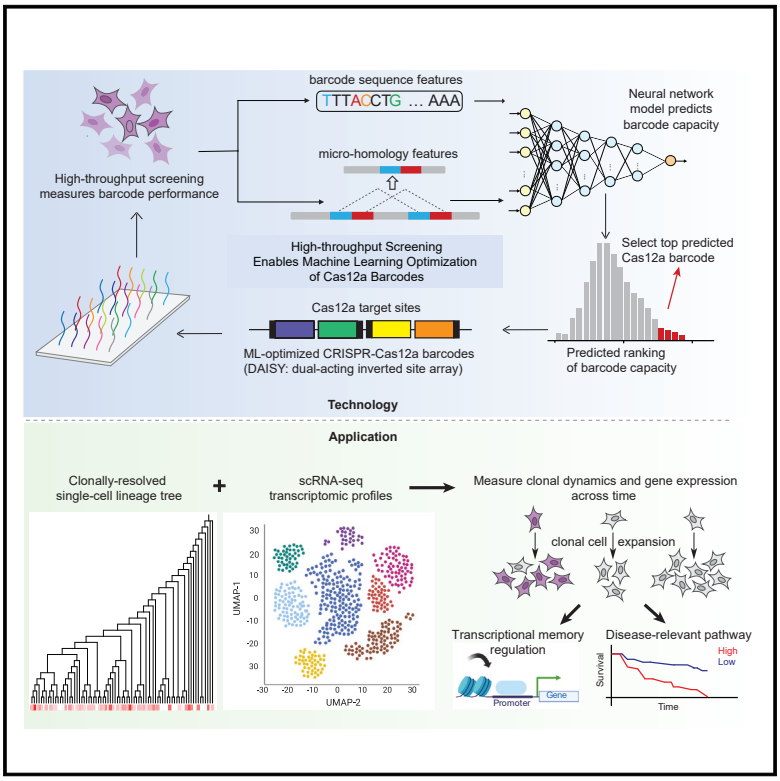
Molecular Cell

Article

Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles

Graphical abstract



Highlights

- DAISY is a CRISPR-Cas12a-based evolvable barcoding technology
- Optimization of DAISY barcodes using an iterative machine-learning pipeline
- Lentiviral delivery enables scalable, tunable barcoding in multiple cancer lines
- Integration with scRNA-seq reveals cell dynamics and transcriptional memory

Authors

Nicholas W. Hughes, Yuanhao Qu, Jiaqi Zhang, ..., Monte M. Winslow, Mengdi Wang, Le Cong

Correspondence

mengdiw@princeton.edu (M.W.), congle@stanford.edu (L.C.)

In brief

Hughes et al. present a single-cell barcoding technology that uses an iterative machine-learning pipeline to optimize the diversity of editing outcomes generated by Cas12a for lineage tracking. Integration with scRNA-seq revealed features of transcriptional memory and nominated EZH2 as a putative epigenetic regulator.

Hughes et al., 2022, Molecular Cell *82*, 1–16
August 18, 2022
https://doi.org/10.1016/j.molcel.2022.06.001

CellPress

Please cite this article in press as: Hughes et al., Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles, Molecular Cell (2022), https://doi.org/10.1016/j.molcel.2022.06.001

Molecular Cell

CellPress
OPEN ACCESS

Article

Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles

Nicholas W. Hughes,[1,2,3] Yuanhao Qu,[1,2,9] Jiaqi Zhang,[5,6,9] Weijing Tang,[1,2,9] Justin Pierce,[1,2,9] Chengkun Wang,[1,2] Aditi Agrawal,[4] Maurizio Morri,[4] Norma Neff,[4] Monte M. Winslow,[1,2] Mengdi Wang,[7,8,*] and Le Cong[1,2,3,10,*]

[1]Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA
[2]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA
[3]Wu Tsai Neuroscience Institute, Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA 94305, USA
[4]Chan Zuckerberg Biohub, Stanford, CA 94305, USA
[5]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[6]Laboratory of Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[7]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544, USA
[8]Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08544, USA
[9]These authors contributed equally
[10]Lead contact
*Correspondence: mengdiw@princeton.edu (M.W.), congle@stanford.edu (L.C.)
https://doi.org/10.1016/j.molcel.2022.06.001

SUMMARY

The development of CRISPR-based barcoding methods creates an exciting opportunity to understand cellular phylogenies. We present a compact, tunable, high-capacity Cas12a barcoding system called dual acting inverted site array (DAISY). We combined high-throughput screening and machine learning to predict and optimize the 60-bp DAISY barcode sequences. After optimization, top-performing barcodes had ~10-fold increased capacity relative to the best random-screened designs and performed reliably across diverse cell types. DAISY barcode arrays generated ~12 bits of entropy and ~66,000 unique barcodes. Thus, DAISY barcodes—at a fraction of the size of Cas9 barcodes—achieved high-capacity barcoding. We coupled DAISY barcoding with single-cell RNA-seq to recover lineages and gene expression profiles from ~47,000 human melanoma cells. A single DAISY barcode recovered up to ~700 lineages from one parental cell. This analysis revealed heritable single-cell gene expression and potential epigenetic modulation of memory gene transcription. Overall, Cas12a DAISY barcoding is an efficient tool for investigating cell-state dynamics.

INTRODUCTION

Uncovering the relationship between cell lineage and gene expression state has contributed to our understanding of the dynamics of multicellular systems (Cao et al., 2020; Neftel et al., 2019; Regev et al., 2017; Tabula Muris Consortium et al., 2018; Tirosh et al., 2016; Travaglini et al., 2020). Mathematical inference has been widely used to estimate cellular trajectories from measurements of the gene expression profiles of cells (Wolf et al., 2019; Setty et al., 2019; La Manno et al., 2018; Saelens et al., 2019; Kester and van Oudenaarden, 2018). These approaches are limited due to the requirement for mathematical assumptions, such as irreversibility of trajectories, which may not match the biological ground truth. An alternative approach is to integrate single-cell transcriptomic profiling with methods that allow direct analysis of cell lineages. Lineage history can be determined by cellular barcoding, where each cell is given diverse DNA sequences as molecular barcodes (Weinreb et al., 2020; Biddy et al., 2018; Rogers et al., 2017; Kebschull and Zador, 2018; Kalhor et al., 2017; Perli et al., 2016). Recently, CRISPR-Cas9 has been utilized to develop evolvable barcoding systems where the editing of a barcode can generate diverse insertions-deletions (indels) that accumulate over time (Alemany et al., 2018; Chan et al., 2019; Raj et al., 2018). In this way, a barcode sequence can evolve into many distinct outcomes (edited sequences, or states), allowing computational reconstruction of subclonal lineages (Jones et al., 2020). The lineage information coupled with single-cell gene expression profiles has allowed interrogation of mammalian (Chan et al., 2019; Bowling et al., 2020) and zebrafish development (Raj et al., 2018), as well as cancer metastasis (Quinn et al., 2021; Simeonov et al., 2021). Although they are sufficient to yield biological insights, existing CRISPR barcodes have not been thoroughly optimized to enable high-capacity lineage tracking. Importantly, unlike endogenous genome targeting, CRISPR barcodes could theoretically use *any* synthetic sequence, a vast space for this design optimization problem.

Please cite this article in press as: Hughes et al., Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles, Molecular Cell (2022), https://doi.org/10.1016/j.molcel.2022.06.001

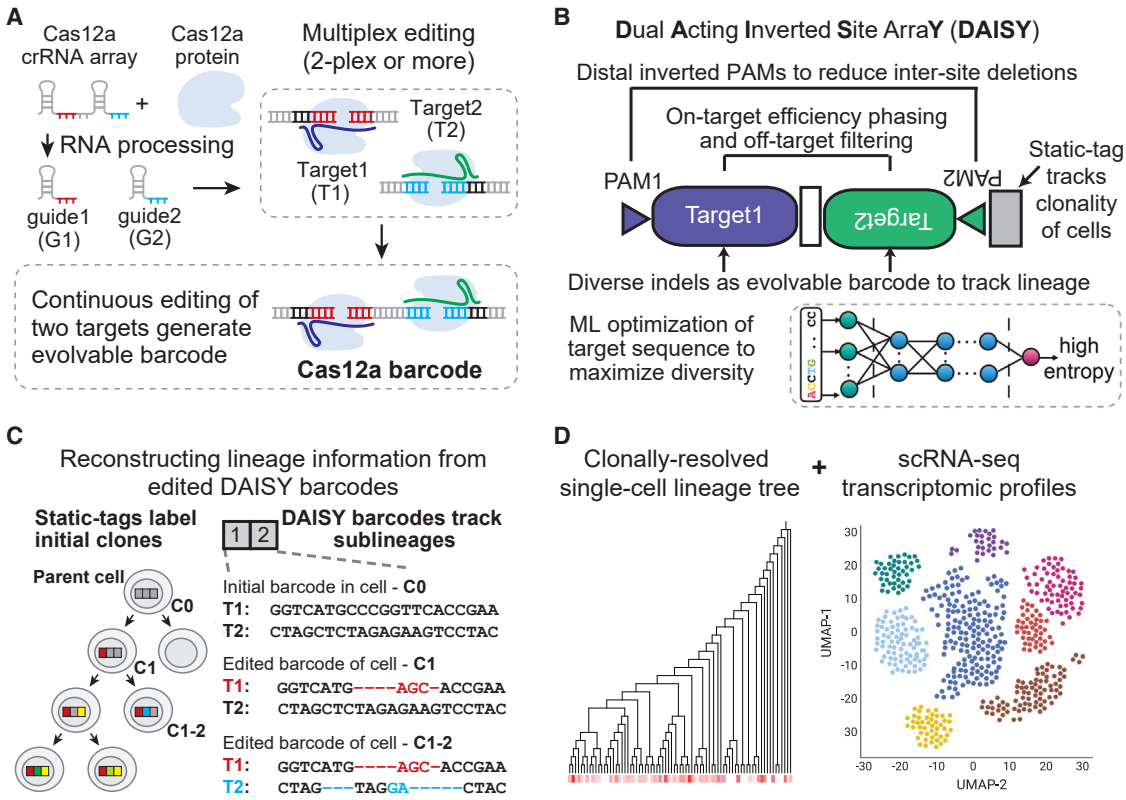


Figure 1. Overview of Cas12a-based DAISY barcodes and pipeline to couple lineage information with single-cell transcriptomic profiling

(A) Design of Cas12a-based barcode system in which a single crRNA array with two guides (G1/G2) could be processed to edit two target sites within a barcode.

(B) Dual acting inverted site array (DAISY) barcode design with two crRNA-target pairs. The guide sequences were selected to have phased editing efficiency (Seq-deepCpf1) and low off-target scores (FlashFry), see STAR Methods for details.

(C) Editing outcomes at the target sites (T1/T2) within a barcode are used to place cells within a lineage tree. Here, an initial edit in T1 allows for the grouping of descendant daughter cells that contain differentiating edits in T2.

(D) Simultaneous recovery of the transcriptome of a cell and an expressed DAISY barcode enables lineage tracking and cell-state classification.

Here, we harnessed the compactness and multi-target editing ability of CRISPR-Cas12a to enable high capacity and tunable molecular barcoding (Figure 1A; Zetsche et al., 2015; Liu et al., 2019). Barcode capacity is the entropy of editing outcomes generated within the barcode. Entropy, measured as the complexity of the evolvable barcode's lineage tree, is a proxy for the number of uniquely trackable lineages (Shannon, 1948). CRISPR barcode entropy directly correlates with its lineage-tracking capacity (Kalhor et al., 2017) as well as the accuracy of recovered lineages (Jones et al., 2020). Barcode tunability is a feature that allows the programmable editing kinetics of a barcode sequence to match the underlying biological process that is being recorded. CRISPR barcodes evolve continuously, and the speed of barcoding determines the time span for recording a biological process, such as lineage commitment. This could be measured via the change rate of the entropy as the barcode evolves, which is analogous to the sampling rate in information theory (Shannon, 1948).

Design

Existing Cas9 barcodes present several challenges. First, there has been no systematic effort to optimize CRISPR barcode sequences to improve lineage-tracking capability. Second, Cas9-based systems present challenges in scalability due to the difficulty of delivering multiple guide RNAs (gRNAs). Compared with Cas9, Cas12a allows a substantially more compact design thanks to the shorter CRISPR RNA (crRNA), and the ability to edit multiple target sites via a single crRNA array. Finally, the Cas12a system has higher targeting specificity than Cas9, which could reduce toxicity in barcoding experiments (Kim et al., 2016).

Current evolvable barcode systems often contain multiple target sites to increase capacity. However, inter-site deletions that span two or more target sites within a barcode are frequent, destroy PAM sequences, and remove a large region of the barcodes (Bowling et al., 2020; McKenna et al., 2016). To overcome this, we designed Cas12a barcodes to have two target sites with phased editing efficiencies to reduce the chance of barcode collapse, which we refer to as dual acting inverted site array (DAISY) barcodes. DAISY features an inverted two-target-site design (Figure 1B), where the Cas12a PAMs are positioned to minimize PAM removal due to inter-site deletion.

Although there have been efforts to predict CRISPR editing outcomes for purposes such as therapeutic gene editing (Chen et al., 2019; Allen et al., 2018; Shen et al., 2018), optimizing a

Please cite this article in press as: Hughes et al., Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles, Molecular Cell (2022), https://doi.org/10.1016/j.molcel.2022.06.001

CellPress
OPEN ACCESS

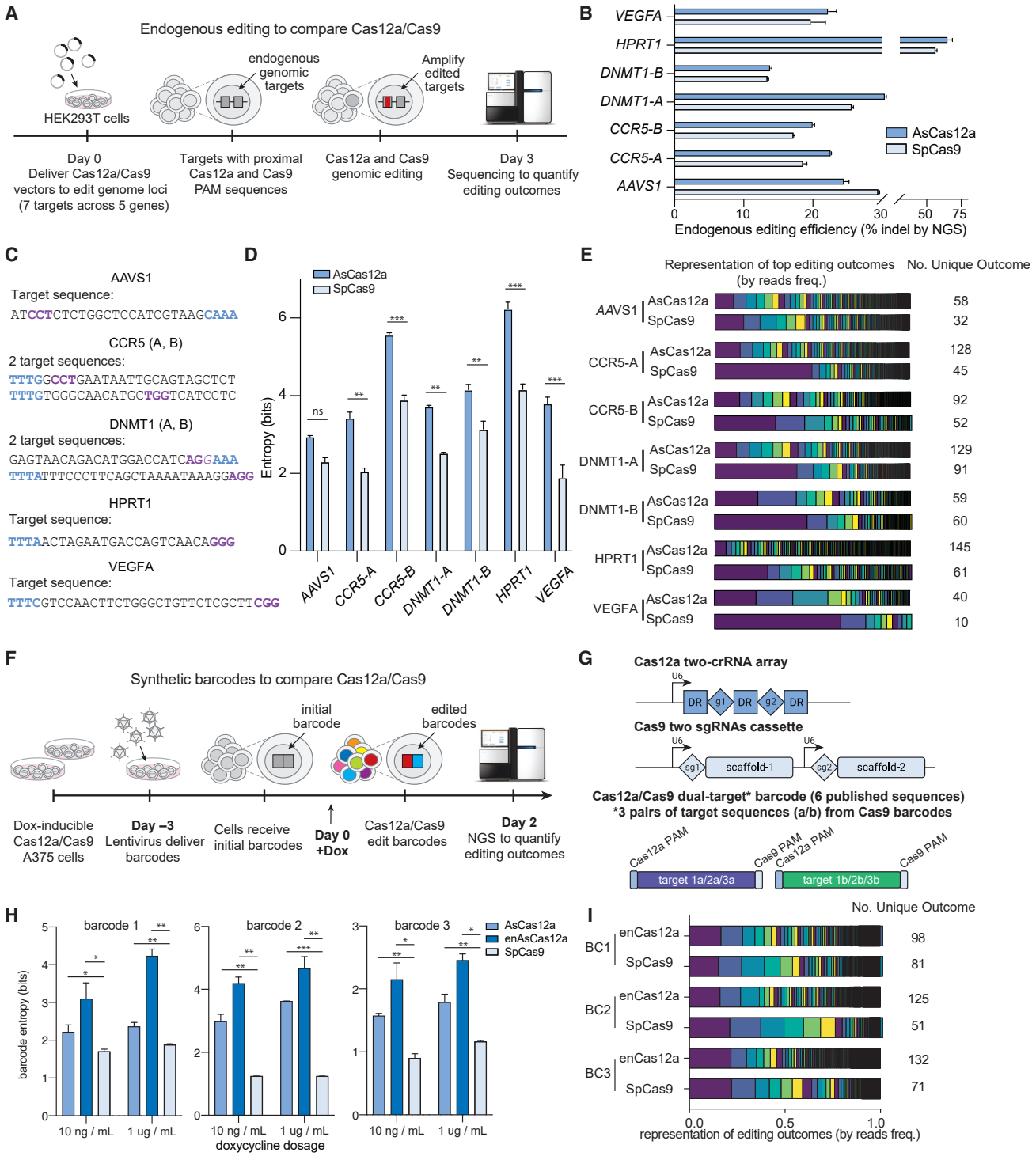


Figure 2. Comparison of Cas9 and Cas12a for gene-editing-based cell barcoding

(A) Design of the endogenous editing experiment to compare Cas12a/Cas9 editing outcomes using transient transfection.

(B) Gene-editing efficiencies across endogenous targets showing comparable levels of indel formation between Cas12a/Cas9.

(C) Endogenous target sequences indicating the proximal PAM sequences (Cas12a in blue, Cas9 in purple).

(D) Entropy of Cas12a- and Cas9-based editing outcomes at endogenous targets.

(E) Stacked bar chart comparing the editing outcome distribution of Cas12a- versus Cas9-based editing outcomes. Bar areas correspond to the sequencing reads frequency of each unique indel outcome.

(legend continued on next page)

Please cite this article in press as: Hughes et al., Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles, Molecular Cell (2022), https://doi.org/10.1016/j.molcel.2022.06.001

CRISPR barcode is more difficult due to the vast sequence choices (a 20-bp CRISPR target sequence has 4^{20} or ~1 trillion possibilities) (Barrangou and Doudna, 2016). We coupled high-throughput screening with an iterative machine learning (ML)-guided search (Figure 1B), which we termed CRISPR learning and optimization via variants exploration with regression (CLOVER). CLOVER generated a collection of high-capacity DAISY barcodes with robust performances. We concatenated DAISY barcodes into a DAISY-chain barcode to enable exponentially higher barcode capacity. We also integrated DAISY barcoding with single-cell RNA sequencing (scRNA-seq) and profiled ~47,000 human melanoma cells. We harnessed the joint lineage and gene expression data to investigate transcriptional memory effects in this cancer model (Shaffer et al., 2020).

RESULTS

Developing a Cas12a barcoding tool and benchmarking with Cas9 barcodes

Cas12a, a class II type V CRISPR-Cas enzyme, has dual RNase and DNase activities (Zetsche et al., 2015). Cas12a binds to the ~20 nucleotide scaffold sequences (DR, direct repeat) and processes a crRNA array to generate multiple crRNAs (Zetsche et al., 2017). The processed crRNAs allow Cas12a to edit multiple target sites (Figure 1A). Cas12a was also shown to have higher specificity versus Cas9 (Figure S1). As Cas12a editing occurs, the initial barcode sequence evolves and branches into multiple lineages (Figure 1C). When coupled with scRNA-seq, a Cas12a barcode enables reconstruction of cellular phylogeny and states from readouts of edited barcodes and transcriptomes, respectively (Figure 1D; McKenna et al., 2016).

To compare the entropy of the repair outcomes generated by Cas9 versus those generated by Cas12a, we evaluated the editing at 7 endogenous genomic loci (Figure 2A). We tested Cas9 and Cas12a targets proximal to one another to control for sequence-based biases (Figures 2A and 2C). Across 6 of the 7 loci, Cas12a led to significantly higher entropy and more diverse editing outcomes, while Cas9 often led to outcomes that were dominated by the most frequent indels (Figures 2D, 2E, and S2D). Importantly, the higher entropy was not due to differential editing efficiency between Cas12a and Cas9 (Figure 2B).

To compare the entropy of Cas9 and Cas12a editing within exogenous barcodes, we generated cell lines with doxycycline-inducible Cas12a or Cas9 (Figures S2A–S2C) and compared the resulting entropies across three barcode sequences extracted from a published study (Bowling et al., 2020; Figures 2F and 2G). We evaluated the entropy of editing outcomes in each barcode after induced expression of Cas12a (AsCas12a), enhancedCas12a (enAsCas12a-HF) (Kleinstiver et al., 2019), or Cas9 (SpCas9) (Figures 2F and 2G). Cas12a editing consistently led to higher barcode entropy than that of Cas9 (Figures 2H and 2I). Collectively, our data indicate that Cas12a generates a wider range of indels and thus higher entropy barcodes. These results motivated us to further develop a Cas12a barcoding system.

Cas12a barcodes reduce inter-site deletions, improve barcode capacity, and allow high-throughput screening

Multi-target Cas9 barcode designs are often employed to increase barcode capacity. However, large inter-site deletions can frequently happen within the barcodes (Bowling et al., 2020; McKenna et al., 2016). Large deletions can lead to descendant cells have undistinguishable barcodes, preventing lineage tracking. Out initial test (Figure 2) also confirmed that levels of inter-site deletion correlate with lower barcoding capacity (up to 3-fold reduction; Figure S2E). To overcome this, we redesigned Cas12a barcodes to have two sites with different editing efficiencies, or phased efficiency, to reduce barcode collapse (Jones et al., 2020). We termed this DAISY barcodes (Figures S3A–S3C). Phased efficiency theoretically minimizes the chances of two simultaneous double-stranded breaks (DSBs), which could result in a large deletion through non-homologous end joining (NHEJ) (McKenna et al., 2016).

The compactness of DAISY barcodes at only ~60 bp enables high-throughput design screening. Although prior work has leveraged pooled screening to measure CRISPR editing (Leenay et al., 2019), there have not been large-scale efforts to uncover optimal CRISPR barcode sequences. Using pooled lentiviral screening, we evaluated the capacity of 14,358 unique DAISY barcodes (Figures 3A and S3A–S3C). We first generated a pool of oligos that contained DAISY barcodes, next to a crRNA array to edit the target sites, and a static tag to uniquely identify each initial DAISY barcode sequence (Figure 3A). For the pool of designs, we selected 5,000 random Cas12a target sites, pairwisely assembled all 25 million combinations, and filtered the pairs to prioritize those with phased efficiency (Figures S3A–S3C; STAR Methods). Using pooled readout, we sequenced the DAISY barcodes across multiple time points, aligned them to the original barcode reference, and quantified their entropy (Figure 3A; STAR Methods).

DAISY barcodes have consistent barcoding activities with low inter-site deletions

Accumulated editing within the DAISY barcodes led to barcode entropy increase over time, with the median rising from ~2 bits at day 2 to ~3.5 bits at day 14 (Figure 3B). We confirmed reproducibility of barcode entropy across two biological replicates (Figure 3C). Also, we observed the high variability of entropy across designs (Figures 3B and 3C), consistent with initial sequences influencing the barcode capacity. Notably, we found that the

(F) Design of synthetic barcode experiments to compare Cas12a/Cas9 using lentiviral vectors and doxycycline-inducible cell lines.
(G) Vector designs for Cas12a editing (top) and Cas9 editing (middle) of a common two-target barcode (bottom). We picked 3 published barcodes from a published Cas9 study (Bowling et al., 2020).
(H) Entropy of editing outcomes within each barcode after doxycycline-induced Cas12a/Cas9 expression.
(I) Stacked bar chart comparing editing outcome distribution as in (E). Unless otherwise noted, all statistical comparisons in this and following figures were performed via a t test with 1% false-discovery rate (FDR) using a two-stage step-up method of Benjamini, Krieger, and Yekutieli, *($p < 0.05$); **($p < 0.01$); ***($p < 0.001$).

Please cite this article in press as: Hughes et al., Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles, Molecular Cell (2022), https://doi.org/10.1016/j.molcel.2022.06.001

CellPress
OPEN ACCESS

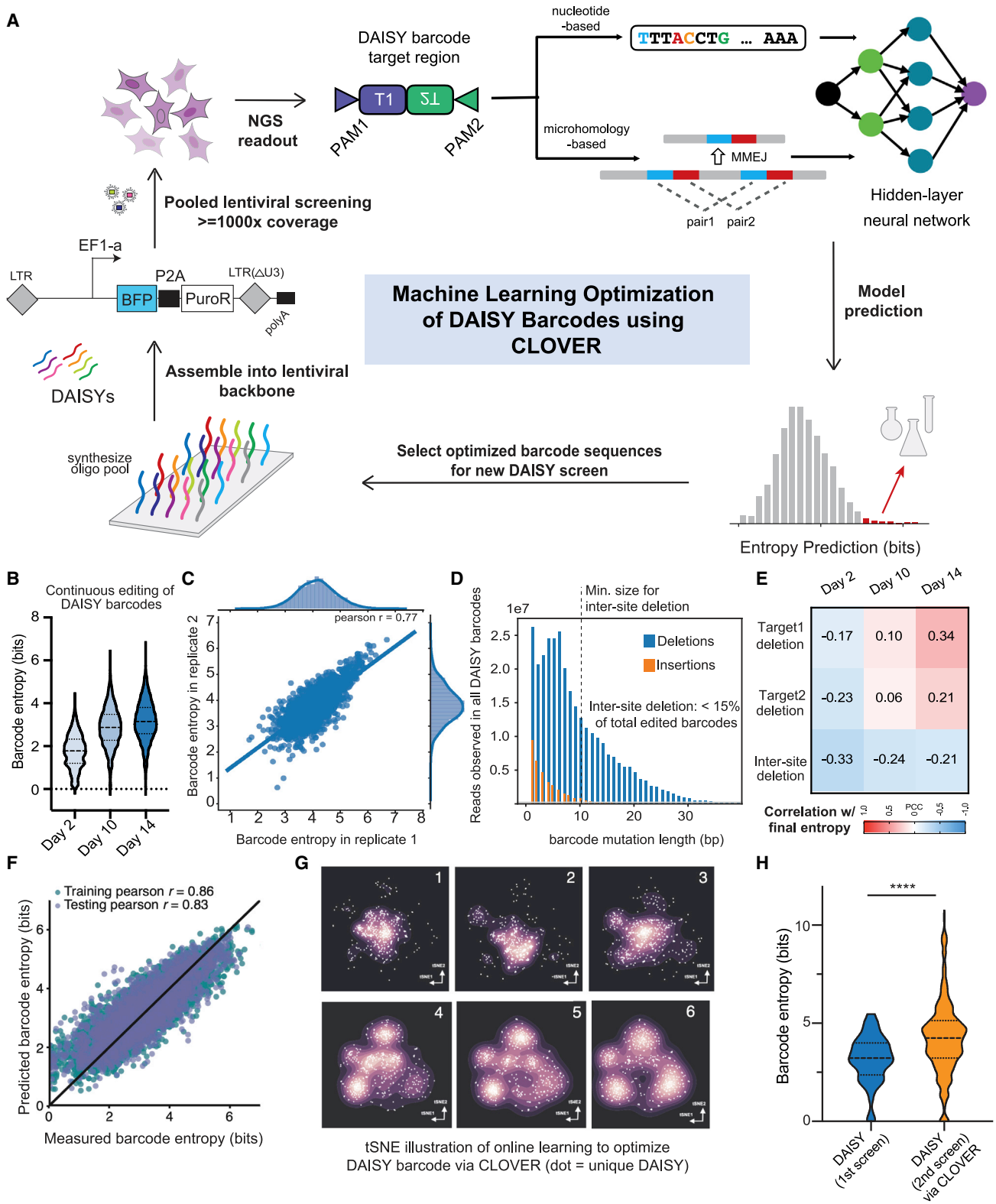


Figure 3. High-throughput screening with ML optimization to generate high-capacity DAISY barcodes

(A) Overall design of CLOVER pipeline to optimize DAISY barcode sequences via iterative pooled screening and ML modeling.

(B) Distribution of barcode entropies across all DAISY barcodes at each time point.

(legend continued on next page)

Please cite this article in press as: Hughes et al., Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles, Molecular Cell (2022), https://doi.org/10.1016/j.molcel.2022.06.001

barcode entropy of DAISY barcodes did not correlate with the Seq-deepCpf1 prediction of editing efficiency (Figure S3D). This indicates that the barcode editing process over two adjacent target sites (a unique feature of the DAISY library) cannot be predicted using existing models based on single-site Cas12a editing data. Furthermore, ~85% of observed indels were fewer than 10 nucleotides in length, providing evidence that DAISY reduces inter-site deletions (any inter-site deletion would be expected to be at least 10 nucleotides in length) (Figure 3D).

Temporal dynamics of barcode editing and their effects on barcode capacity

Next, we analyzed how the temporal dynamics of barcode editing influenced the final barcode capacity. We calculated the pairwise correlation between the three major types of deletions and the measured barcode entropy, across all barcode sequences (Figure 3E). Several trends were immediately apparent. First, early deletion events (on day 2) reduced the chance of further editing and correlated negatively with the final barcode entropy (Figure 3E, day 2 column). Second, there was a strong negative correlation across all time points between inter-site deletion and barcode entropy (Figure 3E, bottom row). In particular, early inter-site deletions at day 2 had the strongest negative correlation. Third, single-site editing at later time points (days 10 and 14) was positively correlated with barcode entropy, which was notable on day 10 and stronger on day 14 (Figure 3E, top two rows). Moreover, we assessed whether Cas12a could retarget a previously edited site, allowing for a continuous increase in barcode entropy. As opposed to Cas9, Cas12a makes PAM-distal cuts that may leave the seed sequence intact, which is required for cleavage (Swarts et al., 2017). We found evidence of disjointed indels that occur within a single-target site, a potential hallmark of retargeting (Figure S3E). The frequency of these relatively rare events increased from ~0.2% to ~3% over time to contribute to the continuous increase in barcode entropy (Figure S3F). Together, these suggest that preventing inter-site deletion and promoting continuous editing are critical to high-capacity Cas12a barcoding.

Machine-learning modeling predicts Cas12a barcode entropy and allows optimization over vast sequence space to generate high-entropy DAISY barcodes

Our initial DAISY library screen suggested that the choice of barcode sequences significantly influences the barcode entropy. Thus, optimal choices of the barcode sequence should maximize lineage-tracking capacity. Exhaustively testing all possible sequences would require the analysis of trillions of possibilities. To address this challenge, we harnessed the predictive power of ML-guided search processes to design an iterative experiment-computation workflow, which we termed CLOVER (Figure 3A). CLOVER uses the data from the DAISY barcode screening to build an ML model that can predict the entropy of untested DAISY barcodes, followed by experimental validation to identify top barcode sequences. The results of the validation can then be added to the data pool to improve the prediction model, thus enabling an iterative pipeline of barcode optimization (Figure 3A; STAR Methods).

Feature selection and model building of CLOVER for optimizing barcode sequence

The CLOVER pipeline consists of three modules: feature engineering, entropy prediction, and path-regularized online learning (Figure 3A; STAR Methods). The first module is a library of features for predictive ML. Inspired by existing models for single-target CRISPR editing (Allen et al., 2018; Chen et al., 2019; Kim et al., 2018; Shen et al., 2018), we constructed a collection of features for the DAISY barcodes, which are based on one-hot encoding of nucleotides, GC content, and a Jaro-Winkler-based distance feature that encoded the process of microhomology-mediated end joining (MMEJ). The Jaro-Winkler gives more weight to the common prefix of two sequences that flank the predicted cut sites. Therefore, it appropriately weighs the increased prevalence of MMEJ-driven editing outcome events as a function of the distance of microhomology tracts from the predicted cut site (Figure 3A).

The second module is for predicting a sequence's entropy. We trained a ridge regression model to test the predictive power of our feature space and found that the model was highly predictive of barcode entropy, with a testing Pearson r of 0.80. To further obtain entropy-guided representations, we trained a neural network using deep learning to arrive at representative features with a testing Pearson r of 0.83, which was used for subsequent modeling and design exploration tasks (Figures 3F and S3G; STAR Methods).

The third module enables adaptive search and dynamic exploration of the design space via *in silico* mutagenesis (Figure 3G; STAR Methods). We developed a path-regularized online learning method using a bandit optimization formulation: at each round of optimization, a learning agent chooses an arm—a combination of designs to experiment on—and receives a stochastic reward. The difference between this reward and the maximal reward at this round, assuming it exists, is defined as instantaneous regret. In the context of our DAISY barcode optimization, the rewards were defined as the average barcode entropy of the identified barcodes, and the instantaneous regret is

(C) Barcode entropy measured at day 14 from two biological replicates, showing consistent results from separate lentiviral transductions.
(D) Indel length distribution across all barcodes where the minimum inter-site deletion length is indicated.
(E) Pearson correlation coefficients (PCCs) between indel outcome types at each time point and the final barcode entropy across all DAISY barcodes.
(F) Neural network model accurately predicts entropy of DAISY barcodes.
(G) 6 rounds of path-regularized online learning were performed (round indicated at top right of each panel). 96 designs are chosen through path regularization (see STAR Methods) in each round (5 simulations total). Therefore, each plot contains 96 × 5 designs, where the kernel density estimation (KDE) is based on the first two tSNE coordinates. The exploration converges on 4 local maxima as indicated by increased point density after 6 rounds.
(H) Distributions of barcode entropy from DAISY barcodes in 1st screen (initial pool) and from 2nd screen (CLOVER-optimized) in A375 cells. *($p < 0.05$); **($p < 0.01$); ***($p < 0.001$), ****($p < 0.0001$).

Please cite this article in press as: Hughes et al., Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles, Molecular Cell (2022), https://doi.org/10.1016/j.molcel.2022.06.001

CellPress
OPEN ACCESS

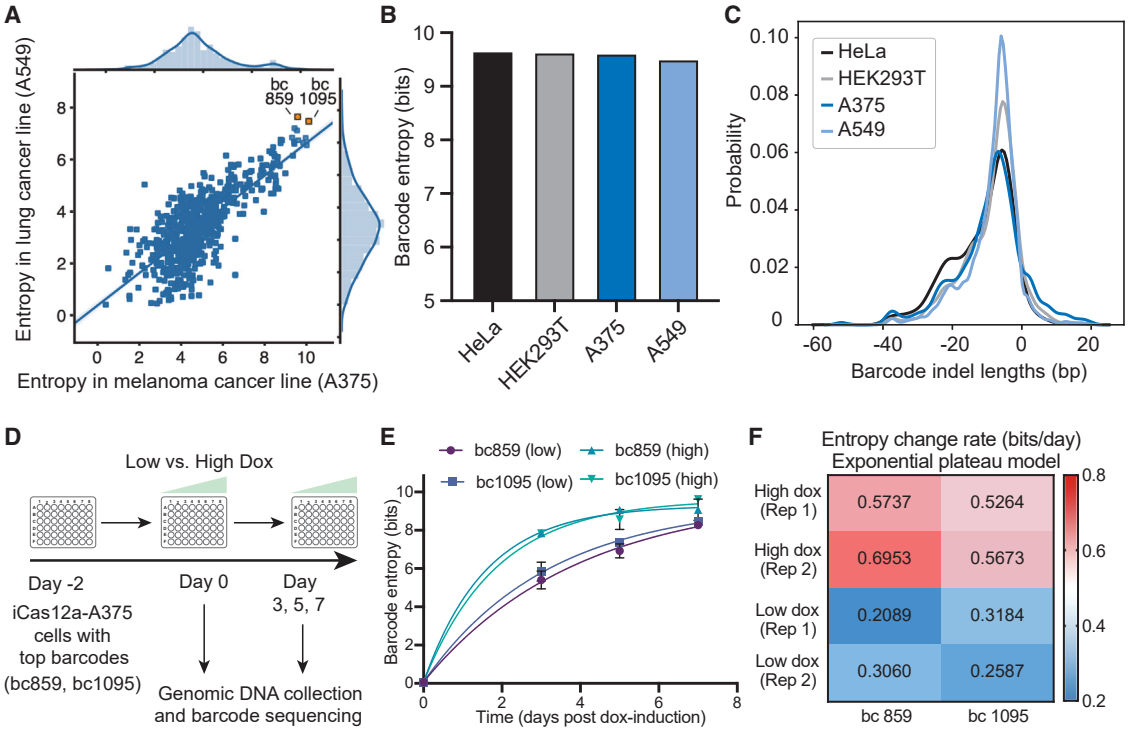


Figure 4. ML-optimized DAISY barcodes have robust performance across cell lines with doxycycline-controllable tunability

(A) Comparison of barcode entropy demonstrating consistent performance of CLOVER-optimized DAISY barcodes in A375 melanoma and A549 lung adenocarcinoma cell lines. Top barcodes used in later experiments are highlighted.

(B) Comparison of total barcode entropy across all clones within each indicated cell type.

(C) Consistent indel mutation length distributions of editing outcomes within the DAISY barcode (BC859) across cell lines.

(D) Experiment design to measure doxycycline-dependent tunability of top DAISY barcodes in A375 cells. Low- and high-dox were 40 and 1,000 ng/mL, respectively.

(E) Change in the barcode entropy over time using low- and high-dox.

(F) Rate kinetics of barcode entropy (based on an exponential plateau model) across doxycycline dosages and biological replicates.

the difference between this value and its maximal-possible value in the pool of sequences. Minimizing the regret is equivalent to maximizing the barcode entropy of sequences that we chose for new experiments. We chose an upper-confidence bandit learning approach for recommending new designs, utilizing a probabilistic surrogate model (Abbasi-yadkori et al., 2011). The approach would recommend random new design sequences with probability proportional to a "potential" score, where the score is a combination of the design's predicted entropy and the prediction's level of uncertainty. This would encourage exploring new designs that are highly dissimilar to tested sequences, which enables fast exploration of large sequence space and quick convergence to the optimal solution (Abbasi-yadkori et al., 2011; Auer, 2002; Rusmevichientong and Tsitsiklis, 2010).

High capacity and tunability of optimized DAISY barcodes

Based on the first DAISY barcode screen, we employed the CLOVER pipeline to generate a new set of 2,000 optimized barcodes to test in a second pooled screen (STAR Methods). We generated a library with optimized barcodes (DAISY 2nd screen) as well as a set of controls from the original screen (DAISY 1st screen) as internal benchmarks (Figure 3H). The ML-optimized DAISY barcodes significantly increased the average barcode entropy compared with the 1st screen, with top barcodes achieving an entropy increase of over 3 bits, suggesting an ~10-fold increase in the number of lineages that the barcode is capable of tracking (Figure 3H; Table S3). We tested the 2nd screen DAISY barcodes in both human melanoma and lung adenocarcinoma cell lines and found that barcode entropies were comparable across cell types. Thus, our second screen validated the exploratory and predictive power of CLOVER to identify optimal barcode sequences within the vast sequence space (Figure 4A).

To further confirm the portability of DAISY barcodes, we tested a top barcode in additional cell types. We delivered this barcode to four different cell lines (A375, A549, HeLa, and HEK293T) derived from lung epithelial, cutaneous skin, cervical, and kidney tissues, respectively. The barcode entropy across cell types was consistently ~9.5 bits after 10 days (Figures 4B and S4A). Further, the indel length distributions showed an enrichment of small deletions (−6 bp) in all lines, suggesting that DNA repair activity within each cell type did not skew the evolution of barcode sequences (Figure 4C). DAISY barcodes thus performed robustly, supporting it as a portable tool for lineage tracking.

Please cite this article in press as: Hughes et al., Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles, Molecular Cell (2022), https://doi.org/10.1016/j.molcel.2022.06.001

Finally, we tested the tunability of two top-performing DAISY barcodes (Figure 4D). Like a recently described one-phase exponential decay model (Park et al., 2021), we derived the entropy change rate of the DAISY barcodes based on our longitudinal measurements (Figures 4E and 4F). We found that by varying the doxycycline inducer concentration, the entropy change rate could be tuned from ~0.25 bits/day (low dosage, slower evolution) to ~0.5 bits/day (high dosage, faster evolution). We further demonstrated the tunability of DAISY barcoding in additional cellular contexts (Figures S4C–S4F). The tunable feature of DAISY barcodes could facilitate applications in which the rate of barcode evolution needs to match the biological processes under investigation (Wagner and Klein, 2020). Taken together, the optimized Cas12a DAISY barcodes are compact, have high capacity, and are tunable.

Combining DAISY barcodes into a high-capacity barcode array (DAISY-chain)

Although the basic DAISY barcodes use a two-target design, many published CRISPR barcodes use 8–10 Cas9 target sites, or deliver more than 20 copies of the same barcode to increase capacity (Bowling et al., 2020; Quinn et al., 2021; Simeonov et al., 2021). Multiple target sites could evolve independently and generate more unique outcomes for lineage tracking. To this end, we concatenated top DAISY barcode sequences (bc859 and bc1095) into a DAISY-chain barcode (Figure 5A). We measured the capacity of the DAISY-chain as in previous experiments (supplemental information). Over the course of 9 days, indels accumulated at the expected cut sites (Figure 5B). Overall, this 120-bp DAISY-chain generated ~66,000 unique edits, reaching over ~12 bits of entropy (Figure 5C). Strikingly, the indel profiles demonstrated the rarity of inter-site deletions (Figure 5D). This contrasts with Cas9 barcode arrays that usually generated large deletions (Bowling et al., 2020; McKenna et al., 2016) and helps to explain its high capacity. Further, we profiled the sequence evolution of the DAISY-chain by assigning alleles to clonal populations using the associated static tag. Over time, most alleles were assigned to a single clone, demonstrating that DAISY barcodes uniquely label subclonal lineages (Figure 5E). Additionally, we designed a competition assay between cells with and without barcode editing to measure genotoxicity from multiplexed Cas12a cleavage. We did not observe competitive advantage of unedited cells, supporting the inference that DAISY barcoding is not genotoxic (Figure S4B). These results support the scalability and low toxicity of the DAISY barcodes.

Cas12a single-cell barcoding profiling recovers lineage and gene expression at scale

To use DAISY barcoding to uncover lineage information and gene expression, we cloned a top DAISY barcode into a lentiviral vector in which edited barcodes would be transcribed and captured by scRNA-seq (Figure 6A). Our single-cell DAISY barcode (scDAISY-seq-v1) vector also contained a static tag to label the founding clonality of cells (supplemental information). We transduced melanoma cells with doxycycline-inducible Cas12a and the scDAISY-seq-v1 vector. We bottlenecked the cells to have approximately 5 parental cells, induced Cas12a to initiate editing of the DAISY barcodes, and harvested cells for scRNA-seq (Figure 6A).

From the single-cell data, we recovered sequencing reads corresponding to the DAISY barcodes in ~2,000 cells, or ~70% of all cells that passed quality filtering (Figures 6A and S5A–S5C; STAR Methods). These single-cell barcode reads harbored a total of 1,512 unique editing outcomes (Figure 6B). The bimodal distribution of the editing events (Figure 6B) demonstrated that most indels were within the target sites. Consistent with our bulk measurements, the DAISY barcode reached an entropy of over 9 bits. We examined the largest clonal population (clone 1, or C1) defined by the static tag, which contained 1,129 cells with 679 unique edited barcodes (Figures 6C and 6D). In C1, 60% of the edited barcodes labeled one descendant cell uniquely (Figure 6D). These edited barcodes from C1 had no observed overlap with those from the second largest clone (Figure 6E). This means that, despite its small size, the DAISY barcode tracked a significant portion of cell lineages at single-cell resolution. This optimized DAISY barcode has a tracking capacity comparable with several Cas9 barcodes, which are often longer or need multiple copies to boost capacity.

Identification of high-memory genes with heritable expression via barcoding

We integrated the lineage history recovered from the DAISY barcode together with single-cell transcriptional profiles to investigate the inheritance of gene expression (Shaffer et al., 2020; Figures 6F and 6G). We calculated the variability of gene expression within DAISY-barcode-defined lineage groups and compared this with a baseline averaged from randomized groups (Figure 6H; supplemental information). We measured the strength of heritable gene expression, or transcriptional memory, by computing a memory index for each gene (Shaffer et al., 2020; Figures 6H and 6I; STAR Methods). Then, by ranking genes according to their memory indices, we identified a subset of high-memory genes (Figure 6I; Table S6).

We examined gene sets enriched within high-memory genes and identified an enrichment for neuronal and chromatin-related pathways (Ashburner et al., 2000; Subramanian et al., 2005; Figure 6J). The association with neuronal genes in melanoma is intriguing, as melanocytes originate from neural crest cells (Zabierowski et al., 2011). Although further investigation will be needed, a rare neural crest stem cell state has been associated with therapeutic resistance in melanoma (Rambow et al., 2018). In addition, the chromatin gene enrichment suggested that a feedback epigenetic mechanism may be involved in maintaining heritable gene expression (Shaffer et al., 2020; Takei et al., 2021). To assess this possibility, we conducted meta-analysis using ENCODE data (ENCODE Project Consortium, 2012) and identified enriched proteins that bound proximally to high-memory genes (Figure 6K). Intriguingly, two top proteins, EZH2 and SUZ12, are members of the polycomb repressive complex 2 (PRC2), which plays key roles in epigenetic regulation (Kim and Roberts, 2016; Holoch et al., 2021; Figure 6K). In support, using chromatin immunoprecipitation sequencing (ChIP-seq) data from melanoma cells (Su et al., 2019), we observed strong enrichment of EZH2 peaks at the transcriptional start sites (TSSs) of high-memory genes, in contrast with control genes with similar expression levels (Figure 6L).

Please cite this article in press as: Hughes et al., Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles, Molecular Cell (2022), https://doi.org/10.1016/j.molcel.2022.06.001

CellPress
OPEN ACCESS

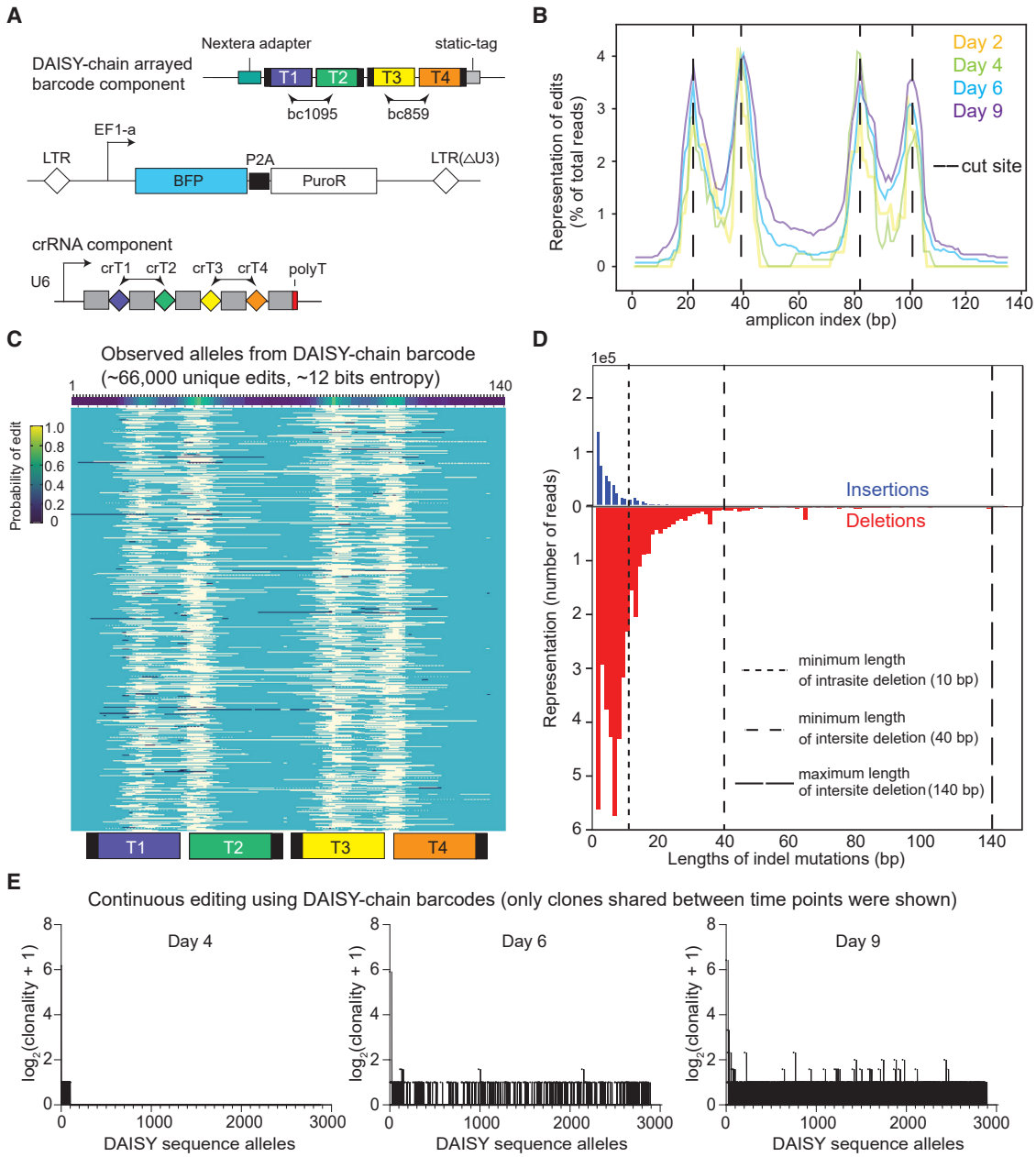


Figure 5. Concatenation of DAISY barcode into a high-capacity DAISY-chain barcode array

(A) Design of a two-DAISY barcode array using top optimized DAISY designs (BC859 and BC1095), encoded in a lentiviral vector.

(B) Editing events distribution within the DAISY barcode array over the 9-day experiment.

(C) Observed barcode alleles generated by the 120-bp DAISY barcode array, with light yellow showing deletions and dark blue showing insertions. The probability of editing derived from all alleles is shown on top and the position of four target sites is shown at the bottom.

(D) Lengths of indel mutations from all alleles using DAISY-chain barcode array. Dash lines marked inter-site deletion limits.

(E) The number of clones associated with each DAISY sequence allele is plotted on the y axis for three different time points (day 4, day 6, and day 9). Each allele is given an index on the x axis.

Investigating clonal dynamics via a time course scDAISY-seq experiment

We next performed a time course scDAISY-seq experiment to characterize clonal dynamics using Cas12a barcoding. We transduced melanoma cells expressing inducible Cas12a with lentivirus containing the DAISY-chain barcode (scDAISY-seq-v2, Figure 7A; supplemental information). We bottlenecked the population such that subclones would be composed of ~10–50 cells by the first analysis at 7 days. Half of the population would then be used to re-seed wells for a final analysis at

Please cite this article in press as: Hughes et al., Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles, Molecular Cell (2022), https://doi.org/10.1016/j.molcel.2022.06.001

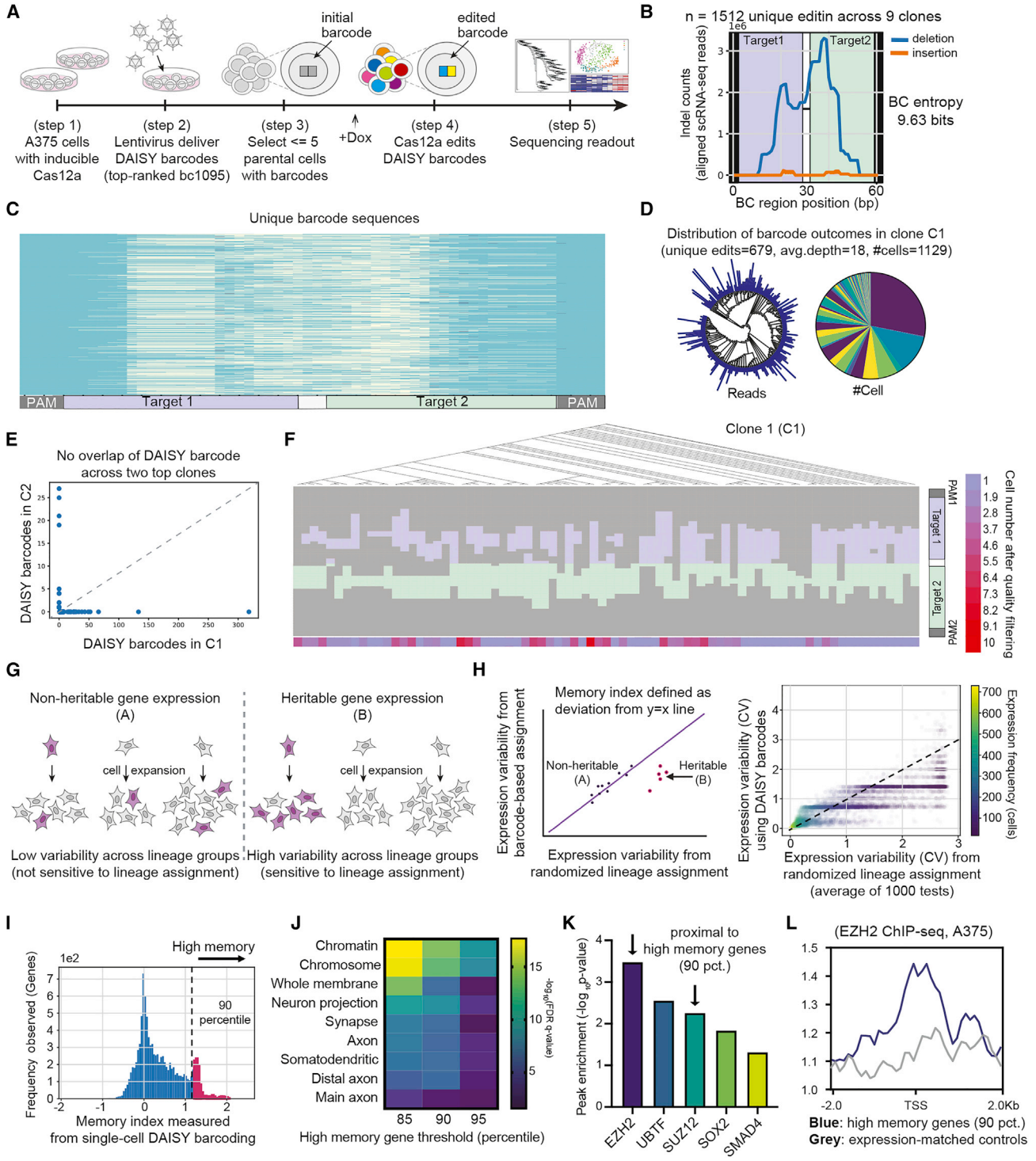


Figure 6. Single-cell demonstration with optimized DAISY barcodes recovers lineage history and transcriptomic information

(A) Design of single-cell experiment using lentiviral delivery of an optimized DAISY barcode (scDAISY-seq).

(B) Distribution of editing outcomes within the DAISY barcode (BC) region. Barcode entropy from single-cell data shown on right.

(C) Unique barcode sequences recovered from scRNA-seq with yellow marks indicating deletions and dark blue marks insertions.

(D) Lineage tree reconstructed from single-cell barcode sequences of largest clone 1 (C1), read counts shown in log scale. Pie charts on the right showing the cell distribution of identified unique lineages.

(E) Homoplasy check showing no overlap between DAISY barcode sequences recovered from the largest two clones, C1 and C2.

(legend continued on next page)

14 days. In total, we recovered 45,914 cells and ~700 clonal populations (Table S5). After filtering (STAR Methods; supplemental information), we recovered 9 clones well-represented across day 7 and day 14. These clones showed stable barcode expression and variable population growth (Figures 7B–7D; Table S5). Cells broadly clustered within transcriptional space, consistent with the expectation that cell-cycle-specific gene expression largely drives cell-state heterogeneity (Figures S5A–S5F).

Across time, editing efficiency of DAISY barcodes increased from a range of 50%–70% to >95% for top-represented clones (C1 and C2) (Figure 7E). To assess genotoxicity of barcoding, we examined the gene expression of cells with varying levels of barcode editing (Figures S5G–S5I). We did not observe evidence of genotoxic effects using a panel of DNA damage response genes (Ihry et al., 2018). After confirming efficient and non-toxic editing, we assessed the diversity of barcode editing within the top clones. We observed minimal overlap between the set of alleles that evolved across clones. Therefore, the barcode capacity of the DAISY-chain was sufficient to prevent homoplasy (Figure 7F). We then analyzed the editing outcomes within a top-represented clone (C1). We observed accumulating barcode indels as the cells continued to proliferate (Figure 7G). These editing outcomes allowed us to perform phylogenetic reconstruction at both time points (Figure 7H). The results showed that the number of subclones increased ~8-fold (from 93 at day 7 to 746 at day 14), while the population size increased by ~12-fold (Figure 7H). These data support the finding that the editing of the DAISY-chain barcode captured lineage bifurcations between the two time points.

Dynamics and reproducibility of transcriptional memory in A375 melanoma cells

We next analyzed data from this longitudinal experiment to evaluate transcriptional memory. We calculated the memory index of each gene and visualized the distribution, which centered at 0 (no memory effect) and had a right skew populated by genes with a putative memory effect (Figures S6A–S6D). We then evaluated the relationship between lineage distance and the memory effect. First, we recalculated the memory index when grouping cousin cells together as opposed to restricting the cell groupings to sister cells that share a most recent common ancestor (MRCA) (STAR Methods). Interestingly, the memory effect was significantly weakened when grouping cousin cells together, as indicated by a shift in the distribution relative to those generated through sister-cell groupings (Figure 7I). The weakened effect gives insight into the permanence of transcriptional memory. If we assume that the barcode was evolving at a rate of ~0.5 bits/day (Figure 4F), and that cousins are differentiable with 2 bits of information (encoding 4 states), then cousins are uniquely marked by a DAISY barcode after ~4 days. Therefore, the memory effect may weaken within this time frame, consistent with previous imaging studies that investigate the dynamics of transcriptional memory and H3K27me3 deposition (Shaffer et al., 2020; Takei et al., 2021).

Next, we analyzed the reproducibility of the memory effects. First, we observed a positive correlation in the memory effect when comparing biological replicates (Figure S6E). Second, we assessed memory effect between clones and found strong evidence for inter-clonal correlation (Figure S6F), which reinforced earlier results from a single clone (Figure 6). Third, we assessed the stability of memory effects over time. For genes with a positive memory index, we observed a positive correlation in memory effects across days 7–14 (Figure 7J). Therefore, the consistency of the observed memory effect across biological replicates, clones, and time points supports the finding that an underlying mechanism regulates transcriptional memory, as opposed to purely stochastic effects (Holoch et al., 2021). Consistent with this and our initial experiment, high-memory genes were found to fall within dedifferentiated cellular states (Figures 7K, S7A, and S7C) and were involved in neuronal functions. Lastly, we again observed evidence of EZH2 binding at the TSSs of high-memory genes at day 7 and day 14 (Figures 7L). Therefore, EZH2 may regulate the cell-state transitions of melanoma cells into dedifferentiated cell states composed of high-memory genes. Taken together, DAISY barcoding coupled with single-cell transcriptomic analysis demonstrated the ability to uncover gene expression dynamics that are otherwise not revealed by static gene expression measurements.

DISCUSSION

We present the Cas12a-based DAISY system as a compact, tunable, and high-capacity CRISPR barcoding method. Our data suggest that Cas12a editing results in intrinsically more diverse editing outcomes than Cas9. This phenomenon may be due to differences in the biochemical properties of Cas12a, with PAM-distal and staggered cutting sites, and its faster dissociation from a genomic locus than Cas9 (Strohkendl et al., 2018; Zetsche et al., 2015; Swarts et al., 2017; Hussmann et al., 2021). We optimized the barcode sequence with an ML pipeline using high-throughput screening data. We provide a detailed list of features used in our CLOVER model, with gradients and potential biological interpretations (Table S3). In addition, our iterative

(F) Reconstructed lineage tree from C1 using DAISY barcodes. Observed edits are illustrated below leaves of the tree. Purple and green bars indicate edits within two target sites. Heatmaps indicate cell numbers after quality filtering.
(G) Illustration of transcriptional memory showing that an expressed gene (amber) can exhibit non-heritable/heritable expression patterns, depending on whether its expression level persists within certain lineages.
(H) (Left) Quantitative definition of a memory index using single-cell transcriptomic data with randomized (x axis) versus barcode-defined (y axis) lineage assignments. (Right) Data from scDAISY-seq were analyzed to calculate memory index for each gene. CV is the coefficient of variation of gene expression (see STAR Methods).
(I) The distribution of memory index values across all genes.
(J) Top significantly enriched gene sets from found high-memory genes.
(K) Top 5 proteins enriched proximally to the high-memory genes (90th percentile) based on ENCODE data.
(L) ChIP-seq peak profiles of high-memory genes (90th percentile) in blue versus control genes (expression-matched, see STAR Methods) in gray.

Please cite this article in press as: Hughes et al., Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles, Molecular Cell (2022), https://doi.org/10.1016/j.molcel.2022.06.001

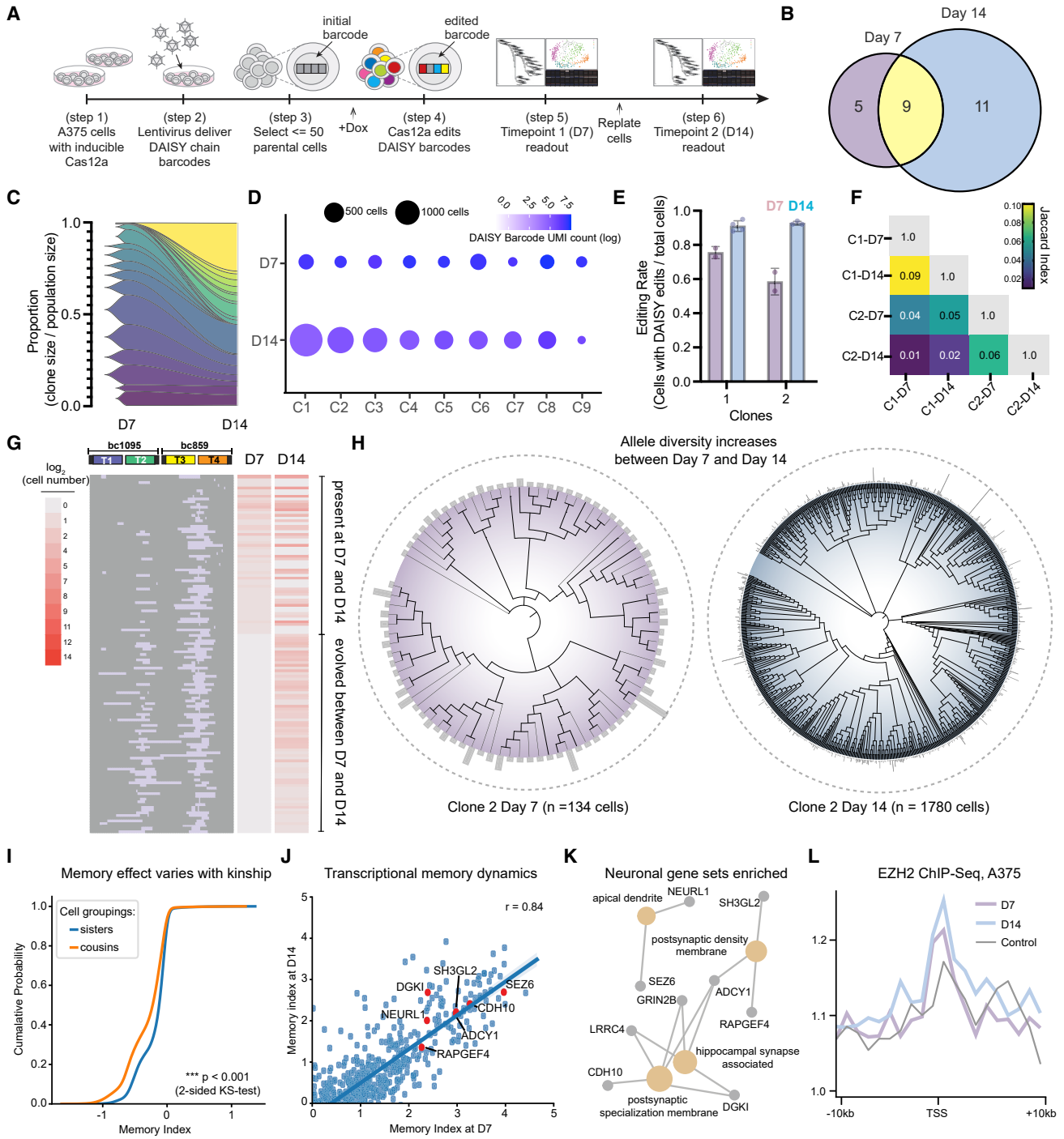


Figure 7. Clonal resampling over time using scDAISY-seq reveals features of transcriptional memory dynamics

(A) Design of the time course scDAISY-seq experiment with clonal resampling. A375 cells expressing inducible AsCas12a were transduced with lentivirus containing DAISY barcodes. Cells were bottlenecked and allowed to proliferate for collections 7 and 14 days post doxycycline induction.

(B) Venn diagram of the resampling of top-ranked clones by population size.

(C) Fish plot of the change in proportions of the top-ranked clone sizes between day 7 and day 14.

(D) Dot plot of the size and expression level across the top-ranked clones.

(E) Measurement of editing rate within two top-represented clones over time.

(F) Sets of alleles within two top-represented clones were compared with each other using the Jaccard index of similarity, where the complete intersection of sets is 1.0 and complete independence of sets is 0.0.

(legend continued on next page)

strategy can be applied to other sequence optimization problems (Yang et al., 2019). Our DAISY-chain barcode array further showcased scalability of the optimized barcodes (supplemental information).

Our scDAISY-seq experiments revealed consistent findings from two independent experiments, including a time course study. Our analysis on heritable gene expression suggested EZH2 (PRC2 complex) as a potential epigenetic regulator of transcriptional memory. Using The Cancer Genome Atlas (TCGA) datasets, we analyzed gene expression of skin cutaneous melanoma (SKCM) tumors from patients and observed that *EZH2* expression levels were positively correlated with the tumor transcriptional heterogeneity, a property that relates to transcriptional memory, and negatively correlated with overall patient survival (Shaffer et al., 2020; Tiffen et al., 2016; Figures S7B and S7D). Future studies using single-cell DAISY-seq could reveal adaptive mechanisms in cancer development, extending beyond the genetically encoded evolution of cancer (Bradner et al., 2017).

CRISPR-Cas9-based single-cell barcoding has been utilized to study a wide breadth of biological processes (Alemany et al., 2018; Chan et al., 2019; Bowling et al., 2020; Spanjaard et al., 2018; Quinn et al., 2021). Our Cas12a barcoding technology will contribute to the expanding lineage-tracking toolkit. Compact DAISY barcodes minimize the likelihood of genotoxic stress due to editing, relative to larger, Cas9-based arrayed barcodes (Meyers et al., 2017; Wang et al., 2017). Given the simplicity of multi-target editing using the Cas12a system, we envision that combining genetic perturbations with scDAISY-seq will be straightforward (Adamson et al., 2016; Datlinger et al., 2017; Dixit et al., 2016). Such a single-cell method to perturb genes and track lineage information will help elucidate how genes control cell fate decisions during development and diseases.

Limitations of the study

Our study introduces a technology for lineage tracking using an ML-optimized Cas12a barcode. The DAISY barcode system could be improved in several ways. For example, additional rounds of CLOVER optimization to uncover better barcodes would, in theory, further increase the tracking capacity. This could help decrease the need for the delivery of multiple target sites. Second, DAISY barcodes could be coupled to functional inputs to record biological signals, such as the activation of signal transduction pathways or cell cycles (Kempton et al., 2020; Tang and Liu, 2018). Finally, further advances in barcoding tunability would allow greater flexibility to record biological processes of different timescales, ranging from the rapid cell differentiation in embryonic development to the gradual process of neurogenesis in the adult brain (Spalding et al., 2013; Ogawa, 1993).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell line and cell culture
- METHOD DETAILS
  - Comparison between Cas9 and Cas12a for molecular barcoding applications
  - DAISY high-throughput screening and machine learning optimization
  - Integration of DAISY barcoding with single-cell RNA-seq (scDAISY-seq)
  - Combination of individual DAISYs into a multi-design array (DAISY-chain)
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Bioinformatic analysis of editing outcomes due to Cas9 and Cas12a nuclease activity
  - Bioinformatic analysis of DAISY barcode editing outcome
  - Feature space design (1st module of CLOVER)
  - Entropy prediction model (2nd module of CLOVER)
  - Path-regularized online learning for barcode optimization (3rd module of CLOVER)
  - Feature extraction to identify motifs and rules defining high entropy DAISY barcodes
  - scDAISY-seq barcode sequencing analysis
  - Cell cycle scoring analysis
  - Transcriptional memory analysis
  - TCGA analysis of EZH2 expression and transcriptional heterogeneity
  - Bioinformatic analysis of editing outcome of DAISY-chain barcodes
  - Sequence evolution analysis of DAISY-chain barcodes

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.molcel.2022.06.001.

(G) Representative profile of indel formation within DAISY-chain barcode from one biological replicate. Indels marked with purple and cell numbers marked with a heatmap.
(H) Phylogenetic reconstructions of a dominant clonal population at day 7 and day 14. Subclonal lineages defined by the DAISY barcode state are at the leaves of the tree and their population sizes are indicated by the adjacent bar heights with a maximum height of 10 cells (left) and 50 cells (right). The height of the bar scales linearly with population size.
(I) Change in the distribution of the memory index within a clone (C2) when grouping cousins together versus sister-cell groupings.
(J) Memory index of genes with positive indices (averaged across all top-represented clones) at day 7 versus day 14 (Pearson correlation coefficient is shown at the top right). A representative group of high-memory genes is highlighted in red.
(K) Gene set enrichment analysis of high-memory genes reveals neuronal gene sets that include dendritic and synaptic biological components.
(L) EZH2 ChIP-seq of high-memory genes across time using genes within the top 85th percentile of the memory index distribution.

ACKNOWLEDGMENTS

We are grateful to members of the Cong and Winslow laboratories and to Jess Hebert, Ravi Dinesh, Sarah Pierce, Feng Pan, Li Zhu, and Will Johnson for support with experiments and discussions on the manuscript. We thank the following scientists: Dr. Feng Zhang (Addgene # 84739) and Dr. Keith Joung (Addgene # 107942). This work was supported by the National Institutes of Health (R35-HG011316 to L.C. and R01-CA231253 to M.M.W.), by Donald and Delia Baxter Foundation (to L.C.), and by National Science Foundation (NSF 1953686 to M.W. and 1953415 to L.C.). This material is based upon work supported by the National Science Foundation Graduate Research Fellowship no. 2018261164 (to N.W.H.) and by Simcere Pharmaceutical Group. The computational analysis is supported by NIH 1S10OD023452 to Stanford Genomics Cluster.

AUTHOR CONTRIBUTIONS

N.W.H., Y.Q., and L.C. designed and performed experiments, analyzed data, and wrote the manuscript. J.Z. and M.W. performed computational analysis, analyzed data, and wrote the manuscript. W.T. performed computational analysis and analyzed data. J.P., C.W., A.A., and M.M. performed experiments. M.M. and N.N. designed experiments and provided reagents. M.M.W., M.W., and L.C. supervised the research and wrote the manuscript.

DECLARATION OF INTERESTS

Stanford University has filed patent applications with L.C. and N.W.H. as inventors on the basis of this work. L.C. is a member of the scientific advisory board of Arbor Biotechnologies.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in science.

Received: November 17, 2021
Revised: April 27, 2022
Accepted: May 29, 2022
Published: June 24, 2022

REFERENCES

Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. Cell *167*, 1867–1882.e21.

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. Adv. Neural Inf. Process. Syst. *11*, 2312–2320.

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

Alexander Wolf, F., Angerer, P., and Theis, F.J. (2018). SCANPY : large-scale single-cell gene expression data analysis. Genome Biol. *19*, 1–5.

Allen, F., Crepaldi, L., Alsinet, C., Strong, A.J., Kleshchevnikov, V., De Angeli, P., Páleníková, P., Khodak, A., Kiselev, V., Kosicki, M., et al. (2018). Predicting the mutations generated by repair of Cas9-induced double-strand breaks. Nat. Biotechnol. https://doi.org/10.1038/nbt.4317.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Auer, P. (2002). Using confidence bounds for exploitation-exploration tradeoffs. J. Mach. Learn. Res. *3*, 397–422.

Barrangou, R., and Doudna, J.A. (2016). Applications of CRISPR technologies in research and beyond. Nat. Biotechnol. *34*, 933–941.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. Nature *483*, 603–607.

Biddy, B.A., Kong, W., Kamimoto, K., Guo, C., Waye, S.E., Sun, T., and Morris, S.A. (2018). Single-cell mapping of lineage and identity in direct reprogramming. Nature *564*, 219–224.

Bowling, S., Sritharan, D., Osorio, F.G., Nguyen, M., Cheung, P., Rodriguez-Fraticelli, A., Patel, S., Yuan, W.-C., Fujiwara, Y., Li, B.E., et al. (2020). An engineered CRISPR-Cas9 mouse line for simultaneous readout of lineage histories and gene expression profiles in single cells. Cell *181*, 1410–1422.e27.

Bradner, J.E., Hnisz, D., and Young, R.A. (2017). Transcriptional addiction in cancer. Cell *168*, 629–643.

Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher-Gonen, R., Zhang, F., et al. (2020). A human cell atlas of fetal gene expression. Science *370*, eaba7721.

Chan, M.M., Smith, Z.D., Grosswendt, S., Kretzmer, H., Norman, T.M., Adamson, B., Jost, M., Quinn, J.J., Yang, D., Jones, M.G., et al. (2019). Molecular recording of mammalian embryogenesis. Nature *570*, 77–82.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics *14*, 128.

Chen, W., McKenna, A., Schreiber, J., Haeussler, M., Yin, Y., Agarwal, V., Noble, W.S., and Shendure, J. (2019). Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. Nucleic Acids Res. *47*, 7989–8003.

Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. Nat. Methods *14*, 297–301.

DeWeirdt, P.C., Sanson, K.R., Sangree, A.K., Hegde, M., Hanna, R.E., Feeley, M.N., Griffith, A.L., Teng, T., Borys, S.M., Strand, C., et al. (2021). Optimization of AsCas12a for combinatorial genetic screens in human cells. Nat. Biotechnol. *39*, 94–104.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell *167*, 1853–1866.e17.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Holoch, D., Wassef, M., Lövkvist, C., Zielinski, D., Aflaki, S., Lombard, B., Héry, T., Loew, D., Howard, M., and Margueron, R. (2021). A *cis*-acting mechanism mediates transcriptional memory at Polycomb target genes in mammals. Nat. Genet. *53*, 1686–1697.

Hussmann, J.A., Ling, J., Ravisankar, P., Yan, J., Cirincione, A., Xu, A., Simpson, D., Yang, D., Bothmer, A., Cotta-Ramusino, C., et al. (2021). Mapping the genetic landscape of DNA double-strand break repair. Cell *184*, 5653–5669.e25.

Ihry, R.J., Worringer, K.A., Salick, M.R., Frias, E., Ho, D., Theriault, K., Kommineni, S., Chen, J., Sondey, M., Ye, C., et al. (2018). p53 inhibits CRISPR–Cas9 engineering in human pluripotent stem cells. Nat. Med. *24*, 939–946.

Jones, M.G., Khodaverdian, A., Quinn, J.J., Chan, M.M., Hussmann, J.A., Wang, R., Xu, C., Weissman, J.S., and Yosef, N. (2020). Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. Genome Biol. *21*, 92.

Joung, J., Konermann, S., Gootenberg, J.S., Abudayyeh, O.O., Platt, R.J., Brigham, M.D., Sanjana, N.E., and Zhang, F. (2017). Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. Nat. Protoc. *12*, 828–863.

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

Kebschull, J.M., and Zador, A.M. (2018). Cellular barcoding: lineage tracing, screening and beyond. Nat. Methods *15*, 871–879.

Kempton, H.R., Goudy, L.E., Love, K.S., and Qi, L.S. (2020). Multiple input sensing and signal integration using a split Cas12a system. Mol. Cell *78*, 184–191.e3.

Kester, L., and van Oudenaarden, A. (2018). Single-cell transcriptomics meets lineage tracing. Cell Stem Cell *23*, 166–179.

Kim, D., Kim, J., Hur, J.K., Been, K.W., Yoon, S.H., and Kim, J.S. (2016). Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. Nat. Biotechnol. *34*, 863–868.

Kim, H.K., Min, S., Song, M., Jung, S., Choi, J.W., Kim, Y., Lee, S., Yoon, S., and Kim, H.H. (2018). Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. Nat. Biotechnol. *36*, 239–241.

Kim, K.H., and Roberts, C.W.M. (2016). Targeting EZH2 in cancer. Nat. Med. *22*, 128–134.

Kleinstiver, B.P., Sousa, A.A., Walton, R.T., Tak, Y.E., Hsu, J.Y., Clement, K., Welch, M.M., Horng, J.E., Malagon-Lopez, J., Scarfò, I., et al. (2019). Engineered CRISPR–Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. Nat. Biotechnol. *37*, 276–282.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498.

Leenay, R.T., Aghazadeh, A., Hiatt, J., Tse, D., Roth, T.L., Apathy, R., Shifrut, E., Hultquist, J.F., Krogan, N., Wu, Z., et al. (2019). Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells. Nat. Biotechnol. *37*, 1034–1037.

Liu, J., Srinivasan, S., Li, C.Y., Ho, I.L., Rose, J., Shaheen, M., Wang, G., Yao, W., Deem, A., Bristow, C., et al. (2019). Pooled library screening with multiplexed Cpf1 library. Nat. Commun. *10*, 3144.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957–2963.

Mahendran, A., and Vedaldi, A. (2015). Understanding deep image representations by inverting them. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE).

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. *17*, 10.

McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science *353*, aaf7907.

McKenna, A., and Shendure, J. (2018). FlashFry: a fast and flexible tool for large-scale CRISPR target design. BMC Biol. *16*, 74.

Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. Nat. Genet. *49*, 1779–1784.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. *34*, 267–273.

Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., et al. (2019). An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell *178*, 835–849.e21.

Ogawa, M. (1993). Differentiation and proliferation of hematopoietic stem cells. Blood *81*, 2844–2853.

Park, J., Lim, J.M., Jung, I., Heo, S.J., Park, J., Chang, Y., Kim, H.K., Jung, D., Yu, J.H., Min, S., et al. (2021). Recording of elapsed time and temporal information about biological events using Cas9. Cell *184*, 1047–1063.e23.

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

Quinn, J.J., Jones, M.G., Okimoto, R.A., Nanjo, S., Chan, M.M., Yosef, N., Bivona, T.G., and Weissman, J.S. (2021). Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. Science *371*, eabc1944.

Raj, B., Wagner, D.E., McKenna, A., Pandey, S., Klein, A.M., Shendure, J., Gagnon, J.A., and Schier, A.F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nat. Biotechnol. *36*, 442–450.

Rambow, F., Rogiers, A., Marin-Bejar, O., Aibar, S., Femel, J., Dewaele, M., Karras, P., Brown, D., Chang, Y.H., Debiec-Rychter, M., et al. (2018). Toward minimal residual disease-directed therapy in melanoma. Cell *174*, 843–855.e19.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The human cell atlas. eLife *6*.

Rice, P., Bleasby, A., and Ison, J. (2000). The EMBOSS users guide (Cambridge University Press).

Rogers, Z.N., McFarland, C.D., Winters, I.P., Naranjo, S., Chuang, C.H., Petrov, D., and Winslow, M.M. (2017). A quantitative and multiplexed approach to uncover the fitness landscape of tumor suppression *in vivo*. Nat. Methods *14*, 737–742.

Rusmevichientong, P., and Tsitsiklis, J.N. (2010). Linearly parameterized bandits. Math. Oper. Res. *35*, 2.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554.

Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., and Pe'er, D. (2019). Characterization of cell fate probabilities in single-cell data with Palantir. Nat. Biotechnol. *37*, 451–460.

Shaffer, S.M., Emert, B.L., Reyes Hueros, R.A., Cote, C., Harmange, G., Schaff, D.L., Sizemore, A.E., Gupte, R., Torre, E., Singh, A., et al. (2020). Memory sequencing reveals heritable single-cell gene expression programs associated with distinct cellular behaviors. Cell *182*, 947–959.e17.

Shannon, C.E. (1948). A mathematical theory of communication. Bell Syst. Tech. J. *27*, 379–423.

Shen, M.W., Arbab, M., Hsu, J.Y., Worstell, D., Culbertson, S.J., Krabbe, O., Cassa, C.A., Liu, D.R., Gifford, D.K., and Sherwood, R.I. (2018). Predictable and precise template-free CRISPR editing of pathogenic variants. Nature *563*, 646–651.

Simeonov, K.P., Byrns, C.N., Clark, M.L., Norgard, R.J., Martin, B., Stanger, B.Z., Shendure, J., McKenna, A., and Lengner, C.J. (2021). Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. Cancer Cell *39*, 1150–1162.e9.

Spalding, K.L., Bergmann, O., Alkass, K., Bernard, S., Salehpour, M., Huttner, H.B., Boström, E., Westerlund, I., Vial, C., Buchholz, B.A., et al. (2013). Dynamics of hippocampal neurogenesis in adult humans. Cell *153*, 1219–1227.

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., and Junker, J.P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. Nat. Biotechnol. *36*, 469–473.

Strohkendl, I., Saifuddin, F.A., Rybarski, J.R., Finkelstein, I.J., and Russell, R. (2018). Kinetic basis for DNA target specificity of CRISPR-Cas12a. Mol. Cell *71*, 816–824.e3.

Su, D., Wang, W., Hou, Y., Wang, L., Yi, X., Cao, C., Wang, Y., Gao, H., Wang, Y., Yang, C., et al. (2021). Bimodal regulation of the PRC2 complex by USP7 underlies tumorigenesis. Nucleic Acids Res. *49*, 4421–4440.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Swarts, D.C., van der Oost, J., and Jinek, M. (2017). Structural basis for guide RNA processing and seed-dependent DNA targeting by CRISPR-Cas12a. Mol. Cell *66*, 221–233.e4.

Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature *562*, 367–372.

Takei, Y., Yun, J., Zheng, S., Ollikainen, N., Pierson, N., White, J., Shah, S., Thomassie, J., Suo, S., Eng, C.-H.L., et al. (2021). Integrated spatial genomics reveals global architecture of single nuclei. Nature *590*, 344–350.

Tang, W., and Liu, D.R. (2018). Rewritable multi-event analog recording in bacterial and mammalian cells. Science *360*, eaap8992.

Tiffen, J., Wilson, S., Gallagher, S.J., Hersey, P., and Filipp, F.V. (2016). Somatic copy number amplification and hyperactivating somatic mutations of EZH2 correlate with DNA methylation and drive epigenetic silencing of genes involved in tumor suppression and immune responses in melanoma. Neoplasia *18*, 121–132.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196.

Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R.V., Chang, S., Conley, S.D., Mori, Y., Seita, J., et al. (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature *587*, 619–625.

Wagner, D.E., and Klein, A.M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. Nat. Rev. Genet. *21*, 410–427.

Wang, T., Yu, H., Hughes, N.W., Liu, B., Kendirli, A., Klein, K., Chen, W.W., Lander, E.S., and Sabatini, D.M. (2017). Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic Ras. Cell *168*, 890–903.e15.

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science *367*.

Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. *20*, 59.

Yang, K.K., Wu, Z., and Arnold, F.H. (2019). Machine-learning-guided directed evolution for protein engineering. Nat. Methods *16*, 687–694.

Zabierowski, S.E., Baubet, V., Himes, B., Li, L., Fukunaga-Kalabis, M., Patel, S., McDaid, R., Guerra, M., Gimotty, P., Dahmane, N., et al. (2011). Direct reprogramming of melanocytes to neural crest stem-like cells by one defined factor. Stem Cells *29*, 1752–1762.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., et al. (2015). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell *163*, 759–771.

Zetsche, B., Heidenreich, M., Mohanraju, P., Fedorova, I., Kneppers, J., DeGennaro, E.M., Winblad, N., Choudhury, S.R., Abudayyeh, O.O., Gootenberg, J.S., et al. (2017). Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. Nat. Biotechnol. *35*, 31–34.

Please cite this article in press as: Hughes et al., Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles, Molecular Cell (2022), https://doi.org/10.1016/j.molcel.2022.06.001

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Endura Electrocompetent Cells	Lucigen	60240-1
One Shot™ Stbl3™	Thermo Fisher	C737303
Chemicals, peptides, and recombinant proteins		
Doxycycline Hydrochloride	Sigma-Aldrich	D3072-1ML
Blasticidin S HCl	Thermo Fisher	A1113903
Puromycin Dihydrochloride	Thermo Fisher	A1113803
Critical commercial assays		
Chromium Next GEM Single Cell 3' GEM, Library & Gel Bead Kit v3.1	10X Genomics Inc	NC1690752
MiSeq Reagent Kit	Illumina	MS-102-2002
Deposited data		
Raw and analyzed data	This study	PRJNA842099
Experimental models: Cell lines		
HEK-293T	ATCC	CRL-1573
HeLa	ATCC	CCL-2
A375	Laboratory of Dr. Paul Khavari	NA
A549	ATCC	CCL-185
Oligonucleotides		
Oligo pool for DAISY screen	Twist Biosciences	N/A
Oligo pool for CLOVER screen	Twist Biosciences	N/A
gBlock for synthetic barcode test	IDT	N/A
Recombinant DNA		
pLenti_enCas12a-HF-STOP_BlastR-EFS-Tet3G-Blast	This study	N/A
pLenti_spCas9-HF-STOP_BlastR-EFS-Tet3G-Blast	This study	N/A
pLenti_AsCas12a-t2a-mKate-EFS-rtTA3-p2a-Puro	This study	N/A
pLenti-U6-cr1-cr2-ampID-T1-T2-polyT-EF1A-BlastR-WPRE for DAISY and CLOVER screens	This study	N/A
pLenti-U6-d1095/d859-capseq2rc(10X Genomics)-polyT-EF1A-BlastR-WPRE	This study	N/A
pY108 (lenti-AsCpf1)	Laboratory of Dr. Feng Zhang	N/A
pCAG-enAsCas12a-HF1	Laboratory of Dr. Keith Joung	N/A
Software and algorithms		
CellRanger	10X Genomics	N/A
EMBOSS	Rice et al., 2000	N/A
FLASh	Magoč et al., 2011	N/A
Cutadapt	Martin, 2011	N/A
CLOVER Pipeline	This study	https://zenodo.org/badge/latestdoi/405663098
deepTools2	Ramirez et al., 2016	N/A
CASSIOPEIA	Jones et al., 2020	N/A
Transcriptional memory analysis	This study	https://zenodo.org/badge/latestdoi/305167447

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw sequencing data	SRA (NCBI)	SAMN28647686
Other		
Detailed protocol for single-cell Cas12a barcoding	This study	Methods S1

RESOURCE AVAILABILITY

Lead contact

Further information and requests for reagents can be directed to and will be fulfilled by the lead contact Le Cong (congle@stanford.edu).

Materials availability

Plasmids will be deposited to Addgene. Primers, cell lines, and any other research reagents generated by the authors will be distributed upon request to other research investigators under a material transfer agreement.

Data and code availability

- Genomic sequencing data have been deposited at the Sequence Read Archive (SRA) and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell line and cell culture

HEK-293T, HEK-293FT, HeLa, HepG2, A375, A549 cell lines were cultured in DMEM, high glucose, GlutaMAX (Gibco) supplemented with 10% fetal bovine serum (FBS, Gemini Bio), 100 U/ml penicillin and 100ug/ml streptomycin (Gibco) at 37 °C with 5% CO_2.

METHOD DETAILS

Comparison between Cas9 and Cas12a for molecular barcoding applications

Inducible Cas9 and Cas12a cell line generation

To generate doxycycline-inducible Cas9 and Cas12a expression vectors, the following reactions were performed. First, backbone vectors containing the Tet-On 3G system (Takara Bio) were digested with AgeI (NEB), EcoRI (NEB) and BamHI (NEB), MluI (NEB), respectively. AsCas12a (iCas12a), enCas12a-HF (ienCas12a-HF), and Cas9 (iCas9) were amplified from template plasmids using compatible primers and cloned into the digested backbones with NEBuilder HiFi DNA Assembly (NEB) (Table S1). The Cas protein sequences were verified through primer walking of the CDS. Lentivirus was produced by co-transfecting the assembled lentiviral vectors with VSV-G envelope and Delta-Vpr packaging plasmids into HEK-293T cells using PEI transfection reagent (Sigma-Aldrich). Supernatant was harvested 48 hr and 72 hr after transfection. A375 cells were transduced at high MOI with 8 μg/mL polybrene using a spin-infection at 1200*g for 45 minutes. After 24 hours, cells were selected with 10 μg / mL blasticidin to establish stably expressing cell lines for inducible barcoding.

In addition, we separately cloned a vector in which AsCas12a was linked to mKate2 with a t2a element, using NEBuilder HiFi DNA Assembly (NEB). Downstream of AsCas12a and mKate2, we cloned a separate CDS containing rtTA3 linked to a puromycin resistance cassette with a p2a element. Lentivirus was produced by co-transfecting the assembled lentiviral vector with VSV-G envelope and Delta-Vpr packaging plasmids into HEK-293T cells following same protocol as previously mentioned. To generate a highly inducible A375 cell line, we induced Cas12a expression using 400 ng/mL doxycycline (Sigma-Aldrich) and sorted the top 10% of mKate2-positive cells. As we fused a "2A-mKate" fluorescent reporter protein with the Cas enzymes, we could monitor and validate their inducible expression via imaging (Figures S2A and S2C). These cell lines supported doxycycline-dependent gene-editing at endogenous human genomic locus with minimal leakage (Figure S2B).

Direct comparison of Cas9 and Cas12a endogenous editing

Three genomic loci were identified that contained both Cas9 and Cas12a PAM sequences within the *AAVS1*, *CCR5*, and *DNMT1* genes. Cloning of gRNA/crRNAs was performed with BbsI or Esp3I (NEB) through a Golden Gate assembly approach into either a Cas12a expressing backbone or a Cas9 expressing backbone, respectively (Table S1). Constructs were sequence verified by

Sanger sequencing using a U6 sequencing primer: 5'-GACTATCATATGCTTACCGT-3'. 7×10^4 HEK-293T cells were transfected with the constructs using Lipofectamine 3000 (ThermoFisher Scientific) in 48-well plates. Genomic DNA was extracted from transfected cells 72 hours later using QuickExtract (Lucigen). The targeted loci were then amplified using Phusion Flash High-Fidelity PCR Master Mix (ThermoFisher Scientific) according to the manufacturer's instructions with primers containing Illumina sequencing adapters. Paired-end reads (150 bp) were generated on an Illumina MiSeq platform.

Synthetic barcode design and cloning

We modified a published Cas9 barcode sequence (Bowling et al., 2020) that contains 10 target sites in two ways (Table S1). First, to make the sequence more compact, we assembled three compact barcodes that contained two target sites by pairing adjacent target sites from the original barcode together. Second, we inserted a Cas12a PAM at the 5' position of each target sequence, thereby allowing direct Cas9 and Cas12a editing comparisons within a synthetic barcode locus (Table S2). To clone the Cas9 gRNAs, we ordered a gBlock gene fragment (IDT) containing gRNA1, scaffold sequence, mU6 and gRNA2. The fragment was cloned into a Esp3I digested lentiviral vector containing a hU6 promoter using NEBuilder HiFi DNA Assembly (NEB). The Cas12a crRNA array and target array were cloned through OE-PCR using Phusion Flash High-Fidelity PCR Master Mix (ThermoFisher Scientific) followed by NEBuilder HiFi DNA Assembly (NEB) according to the manufacturer's instructions. Constructs were verified by Sanger sequencing using a U6 sequencing primer and a WPRE sequencing primer. Lentivirus was produced by co-transfecting the library with VSV-G envelope and Delta-Vpr packaging plasmids into HEK-293T cells following same protocol as previously mentioned.

Lentiviral delivery of synthetic barcode

A375 cells stably expressing iCas12a, ienCas12a-HF, and iCas9 were transduced at high MOI with the barcodes as described above. 72 hours after transduction, 1×10^5 cells were collected for gDNA extraction using Quick Extract (Lucigen). Doxycycline was added to the media of the remaining cells at the following doses: 0 ng/ mL, 10 ng / mL, and 1000 ng / mL. Cells were collected 2 days after doxycycline induction for gDNA extraction. Barcodes were amplified with primers containing Illumina adapters using NEB Ultra II Q5 Master Mix (NEB) according to the manufacturer's instructions. Paired-end reads (150bp) were generated on an Illumina MiSeq with a 75k read depth per sample.

DAISY high-throughput screening and machine learning optimization

DAISY barcode library design

We generated a library of 5000 random Cas12a targets, filtered for low off-target activity, GC content, and polyT stretches using FlashFry (McKenna and Shendure, 2018). Briefly, a composite off-target score was created that took into account (1) the total number of off-target sites (otCount), (2) the smallest basepair distance (Bpdiff) between an off- and on-target site, and the number of off-target sites within Bpdiff (closeCount):

$$OT = \ln\left(otCount * \left(\frac{closeCount}{Bpdiff}\right) + 1\right)$$

Target regions were then scored for their predicted indel efficiency for each target region using DeepCpf1 (Kim et al., 2018). A custom script was then used to partition a target region into an efficient editing category (DeepCpf1 score $\geq$ 50) and an inefficient editing category (DeepCpf1 score $<$ 50) and to create all pairwise combinations of efficient and inefficiently targeted sequences. Next, all pairwise combinations were scored for their average efficiency score and standard deviation of efficiency scores to create a composite editing score for the combined pair. The coefficient of variation of the composite editing score was computed and used to rank all pairwise combinations. The top 55th percentile target regions were chosen (14358 unique sequences) to assemble into a final pool of oligonucleotides (Twist Biosciences). As negative controls, 12 barcode sequences in which the spacer and target sequences were mismatched were included. Each DAISY barcode sequence within the 14358 pool was uniquely identified with a 10 bp sequence as static tag (Figures S3A–S3C).

For the second screen using an optimized DAISY barcode library, we utilized the CLOVER model to identify 2000 barcode sequences predicted to have increased barcode diversity. Briefly, seed barcode sequences were iteratively mutated to explore the diverse sequence space. 10 iterations were performed to identify the final set of 2000 barcode sequences. The 60 nucleotide barcode sequences were then synthesized along with their paired crRNAs and a unique 10 bp NGS tag sequence as above.

DAISY barcode library cloning

The lentiviral vector was constructed using NEBuilder HiFi DNA Assembly (NEB) to remove the existing Cas9 scaffold sequence and to incorporate two Esp3I (NEB) restriction sites downstream of an AsCas12a direct repeat sequence (Figure 3A; Table S1). Then, it was digested with Esp3I (NEB) at 37oC for 1 h and gel-purified with Monarch DNA Gel Extraction Kit (NEB). Oligonucleotides from Twist Bioscience, with dual crRNAs paired with their targets, were resuspended to 10 ng/uL. The oligonucleotide pool was PCR amplified using KAPA Biosystems HiFi HotStart ReadyMix (2X) and gel-purified. The amplified library was then assembled into the library backbone using NEBuilder HiFi DNA Assembly (NEB) with a molar ratio of 20:1, respectively. The assembly reaction was then precipitated and resuspended in TE buffer (Macherey-Nagel) according to a previously published protocol (Joung et al., 2017). The entire reaction was used to transform Endura Electrocompetent Cells (Lucigen) following the manufacturer's protocol. The transformed cells were cultured at 25°C for 48 hours to minimize recombination between direct repeats. Colonies were then harvested directly and plasmid DNA was extracted with a Plasmid Plus Maxi Kit (Qiagen).

DAISY barcode next-generation-sequencing library generation

Genomic DNA was extracted with DNeasy Blood & Tissue Kit (Qiagen) following the manufacturer's protocol. A first round PCR reaction was performed using 300 ng of genomic DNA, 0.2 μl 100 mM primer and 25 μl NEBNext Ultra II Q5 Master Mix (NEB) per reaction. A total of 12 reactions were performed per sample to ensure adequate representation of the lentiviral pool. PCR reactions were performed according to the following conditions: 98°C 30 s, followed by 25 cycles of 98°C 10 s, 60°C 20 s, 72°C 20 s, followed by 72°C 2 min. The PCR product was pooled and cleaned with 0.8X CleanNGS DNA SPRI Beads (Bulldog Bio) and resuspended in 20 uL of elution buffer (Table S1). The resulting purified PCR products were then quantified and 10-20 ng were loaded into a second round of amplification using NEBNext Q5 Master Mix (NEB) to incorporate flow-cell adaptor sequences and sample indexes to enable demultiplexing of pooled samples. The PCR reaction was performed with the following conditions: 98°C 30 s, followed by 15 cycles of 98°C 5 s, 72°C 20 s. The resulting libraries were then cleaned with 0.8X CleanNGS DNA SPRI Beads (Bulldog Bio), equimolarly pooled, and then gel-purified. The resulting pooled library was sequenced on an Illumina HiSeq 4000 sequencer using paired-end 150 cycle reads.

Integration of DAISY barcoding with single-cell RNA-seq (scDAISY-seq)

scDAISY-seq vector construction

For the initial scDAISY-seq experiment using a two target (2T) design, a top performing barcode (bc1095) was synthesized to include the following components: (1) the crRNA targeting the evolvable molecular barcode (82 nt) (2) the evolvable molecular barcode targets (60 nt) (3) a static tag (10 nt) and (4) a 10x Genomics capture sequence (22 nt) (Figures S3 and S5; Table S1). These components were cloned downstream of the hU6 promoter through Esp3I digestion (NEB) followed by Gibson assembly using NEBuilder HiFi DNA Assembly (NEB). The 10x Genomics capture sequence (22 nt) allows for binding of the expressed molecular barcode directly to 10x Genomics gel beads contained within the Chromium Next GEM Single Cell 3' Reagent Kit v3.1.

For the time course experiment using the DAISY-chain barcode (d1095/d859 combined), the target region of the barcode was cloned into the UTR of eGFP placed downstream of a Tet-On 3G (Takara) TRE promoter by Gibson assembly using NEBuilder HiFi DNA Assembly (NEB) into a MluI-digested lentiviral backbone. The barcode target cloning fragment was initially constructed using a nested overlap extension (OE) PCR reaction in which an ultramer (IDT) with partial homology to a short oligo bearing a random 10bp sequence and a Nextera (Illumina) Read 2 sequence was performed to generate a dsDNA cloning fragment. The cloning fragment was then used for a low-cycle (10 cycles) PCR using Q5 (NEB) polymerase following the manufacturer's instructions to introduce homology arms for the final Gibson assembly. In parallel, the hU6 and crRNA array targeting the DAISY-chain barcode were PCR amplified from a template plasmid used in bulk benchmarking experiments (Figure 5) and inserted into the NotI site of the lentiviral backbone.

Lentiviral delivery of barcodes for single cell barcoding

Lentivirus was produced by co-transfecting the scDAISY-seq vector with VSV-G envelope and Delta-Vpr packaging plasmids into HEK-293T cells following same protocol as previously mentioned. Then, it was used to transduce at least 5×10^4 A375 cells harboring doxycycline-inducible AsCas12a. Cells were then bottlenecked through limited dilution to contain ~ 5 clones / well in a 96-well plate for the initial experiment and ~50 clones / well for the DAISY-chain time course experiment. More cells were seeded initially for the time course to enable an earlier readout 7 days post doxycycline induction. Upon seeding, Cas12a expression was induced with 400 ng / mL doxycycline. For the initial experiment, cells were then expanded for approximately two weeks for barcode editing. For the time course experiment, cells were collected 7 days post-induction and a fraction (~50%) of the cells were replated for a second collection at 14 days post-induction. At each collection, cells were harvested for single-cell RNA-sequencing using the Chromium Next GEM Single Cell 3' Reagent Kit v3.1 under the manufacturer's protocol unless otherwise noted.

scDAISY-seq barcode sequencing using the 10X capture sequence

The cDNA library from step 4.2 of the manufacturer's protocol (10x genomics), was used as a template for PCR. A forward primer, binding to the Nextera (Illumina) Read 1 sequence, was used with a reverse primer that binds specifically to the expressed barcode sequence at the terminal direct repeat (Table S1). The PCR was performed using NEB Ultra II Q5 Polymerase following the manufacturer's protocol. Barcode libraries were sequenced with paired-end 150 cycle configuration on a MiSeq instrument (Illumina).

scDAISY-seq barcode sequencing using a polyA capture approach

The cDNA library from step 4.2 of the manufacturer's protocol (10x Genomics), was used as a template for PCR. A forward primer, binding to the Nextera (Illumina) Read 1 sequence, was used with a reverse primer that binds specifically to a Nextera (Illumina) Read 2 sequence that was cloned next to the DAISY chain target sequence (Table S1). The PCR was performed using NEB Ultra II Q5 Polymerase following the manufacturer's protocol. Barcode libraries were sequenced with a paired-end 150 cycle configuration on a NextSeq 500 instrument (Illumina).

Combination of individual DAISYs into a multi-design array (DAISY-chain)

DAISY-chain vector construction

The lentiviral backbone was first cloned to encode a crRNA array targeting the top two DAISY barcodes (bc1095 and bc859) upstream of the WPRE using a Golden Gate ligation protocol (Table S1). Then, custom oligos were used to generate Klenow fragments

that contained the DAISY target sequences and diverse static clonal tag. The Klenow fragments were PCR-amplified using Q5 (NEB) Polymerase for 10 cycles and assembled into the lentivirus backbone. The product library was transformed at library-scale using the same protocol as the DAISY screening library.

Cell line generations and lentiviral delivery of DAISY-chain vector

Lentivirus, containing DAISY-chain/bc1095/bc859 barcode and a diverse static tag, was generated using the protocol as previously mentioned. In parallel, A375, A549, HeLa, HEK-293FT, and HepG2 cells were transduced using either a constitutive enCas12a (DeWeirdt et al., 2021) or an inducible enCas12a-HF lentiviral construct to generate stable cell lines with constitutive (consCas12a) or inducible enCas12-HF (iCas12a) expression, respectively. The resulting consCas12a and iCas12a cells were transduced with the DAISY-chain lentivirus, bc859 lentivirus, or bc1095 lentivirus at low MOI (~0.3) such that each cell, on average, contained a single static tag. Cells containing integrated DAISY barcodes were then selected with puromycin and allowed to expand during a time course in which gDNA was extracted using QE (Lucigen). For the doxycycline tunability experiments, cells were collected prior to induction and subsequently cultured in either 10 ng / mL doxycycline or 100 ng / mL doxycycline.

NGS Library generation and sequencing

Approximately 1×10^4 cells were used to amplify the target sites within the DAISY barcode using a nested PCR approach with NEB Ultra II Q5 Master Mix (NEB). Briefly, the target sites were initially amplified for 30 cycles. The resulting PCR products were SPRI bead-purified (0.7X ratio) and then loaded into a second-round low-cycle (5 cycles) PCR to introduce Illumina adaptors and multiplexing indexes for sample pooling. The resulting libraries were pooled, column-purified and sequenced on an Illumina MiSeq using a 2x150 configuration.

QUANTIFICATION AND STATISTICAL ANALYSIS

Bioinformatic analysis of editing outcomes due to Cas9 and Cas12a nuclease activity

Endogenous target and synthetic barcode sequencing data were analyzed using a custom pipeline. Briefly, reads were split into reads 1 and 2 and then merged using flash v1.2.11 with default parameters (Magoč and Salzberg, 2011). Next, for barcoding data, we demultiplexed the barcodes by computing the minimum Levenshtein distance between the read sequence and the reference sequence. Then, for both endogenous target and synthetic barcode, we aligned the reads to their corresponding reference using *needleall* (Rice et al.). Finally, the information content contained within the target sites was computed as the Shannon entropy:

$$H(X) = -\sum_{i=1}^{n} P(x_i) log P(x_i)$$

Where x_i is a unique editing outcome generated by either Cas9 or Cas12a (Tables S2 and S4). Relevant statistical tests and parameters used can be found in the figure legends (Figure 2).

Bioinformatic analysis of DAISY barcode editing outcome

We first demultiplexed these paired-end reads by their 10 bp amplicon barcode sequence. We then parallel aligned these reads to their amplicon barcode-assigned reference sequence using *needleall* with the following scoring matrix: match = 5, mismatch = -4, gap-open = -20, gap-extension = -0.5, where mismatch penalties for *N*s, Vs and *B*s were set to 0 (Rice et al.). The generated indel profiles for the sample collected at Day 0 (day of doxycycline-induction) were used to filter out indels observed in samples collected at later time points. To enable comparison of barcodes with variable read depth, we down-sampled barcodes such that all barcodes were uniformly represented with 500 reads. The resulting indel profiles were used to define the mutational outcomes of Cas12a nuclease activity using the Shannon entropy (barcode diversity) as previously described. Relevant statistical tests and parameters used can be found in the figure legends (Figure 3).

Feature space design (1st module of CLOVER)

As input to our predictive model in the CLOVER pipeline, representative features are necessary to capture the characters of DAISY barcodes that contribute to difference in editing outcomes and thus in entropy. We designed a 4906-dimensional feature space that includes both nucleotide-based and microhomology-based information (Figure 3A). We used one-hot encoding to include 54×4-single-nucleotide features and 53×16-dinucleotide features in our nucleotide-based features concerned with the 60bp target region (common base pairs in PAMs not included). Since spacers matched the targets, we added the varying lengths (21, 23 or 25 nt) of the two guides from spacers as 2-length-dependent features. In consideration of the fact that the base pairs closest to the cutting site take on a comparatively heavier responsibility for Microhomology Mediated End Joining (MMEJ)-based deletion, we used the Jaro-Winkler distance, which gives more weight to the common prefix of two sequences that flank a deleted region within the barcode. We calculated the proportion of GC-base pairs in these two subsequences to account for GC contents as another feature. There is a total possibility of 53×52/2 deletions (base pairs in PAMs not considered). For each deletion, we consider the two 15-bp (or less, depending on the length of the deletion as well as if there are enough base pairs flanking it)-subsequence-pairs flanking both the left and right sides of the deletion. This leads to a total of 1920×2-microhomology-based features.

Please cite this article in press as: Hughes et al., Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles, Molecular Cell (2022), https://doi.org/10.1016/j.molcel.2022.06.001

Entropy prediction model (2nd module of CLOVER)

The full barcode dataset was divided into a 70%-training set and a 30%-testing set. Our first predictive model utilized Principal Components Analysis on the training set to find directions of the feature space with high variances. On top of the 256 principal components with highest explained variances, a Ridge regression model was trained to predict entropy. Let X be the input of principal components and Y be the output of entropy, ridge regression seeks to minimize the following objective with respect to θ

$$\|Y - X\theta\|_2^2 + \alpha\|\theta\|_2^2.$$

We used a 5-fold cross validation to pick the l_2-penalty α. This model achieved a training Pearson $r = 0.83$ and a testing Pearson $r = 0.82$. Next, we used deep learning to train a second model directly on top of the original feature space, which aimed to find entropy-representative features (instead of features with high variances given by PCA). We found that a four-hidden-layer fully-connected neural network with ReLU activation (Figure S3G) was able to achieve comparable results with training Pearson $r = 0.86$ and testing Pearson $r = 0.83$.

Path-regularized online learning for barcode optimization (3rd module of CLOVER)

To optimize for barcode design, we developed a path-regularized online learning method using a bandit formulation: in round t of the experiment, the agent picks an arm X_t (a 60bp target region) from a given decision set $\Omega_t \subset \{A, T, G, C\}^d$ (feasible target designs where d = 60) for testing. Subsequently, they receive a reward Y_t (barcode entropy), such that

$$Y_t = f(X_t)^\top \theta_* + \eta_t,$$

is modeled as a linear transform of $f(X_t)$. Here $f : \{A, T, G, C\}^d \to \mathbb{R}^n$ is a transform function that maps an arm to a representation for reward prediction and is inherited from the second module (256 PCs in the first model or outputs from the second-to-last layer in the second model). $\theta_* \in \mathbb{R}^n$ is an unknown parameter and $\eta_t \in \mathbb{R}^n$ is a random noise satisfying conditional zero mean and sub-Gaussian ($\mathbb{R}$ is the set of all real numbers and $n = 256$ in both models). To maximize the instantaneous reward, the agent seeks

$$x^* = \text{argmax}_{x \in \Omega_t} f(x)^\top \theta^*.$$

Since θ_* is unknown, one has to estimate it, however best estimation based on current information might stress too much on exploitation and avoid exploring unknown arms. To address this trade-off, we use the upper confidence bound method[6-8] by constructing a confidence ellipsoid for θ_*. Denote $f(X_{1:t}) = (f(X_1)^\top, \ldots, f(X_t)^\top)^\top$ and $Y_{1:t} = (Y_1, \ldots, Y_t)^\top$ as all past collected data. Let $\widehat{\theta}_t$ to be the ridge estimator of θ_* using these data, i.e.,

$$\widehat{\theta}_t = \text{argmax}_\theta \|Y_{1:t} - f(X_{1:t})\theta\|_2^2 + \lambda\|\theta\|_2^2$$

Under some assumptions, we have, with high probability, θ_* that lies in a ellipsoid centering at $\widehat{\theta}_t$

$$C_t = \left\{\theta \in \mathbb{R}^n : \|\theta - \widehat{\theta}_t\|_{V_t} \le c\right\},$$

where $V_t = \lambda I + f(X_{1:t})^T f(X_{1:t})$, $\|\widehat{\theta}_t - \theta\|_{V_t} = (\widehat{\theta}_t - \theta)^T V_t (\widehat{\theta}_t - \theta)$ and c is some constant. The UCB method then picks the arm for next round of experiments by

$$x^* = \text{argmax}_{x \in \Omega_t,\ \theta \in C_t} f(x)^\top \theta.$$

To enable fast exploration of large sequence space of DAISY designs, we introduced a path-regularized search for the feasible target space Ω_t. This target space is constructed in an iterative adaptive manner. At each round t, we first calculate a hypothesis set of arms from Ω_{t-1} by maximizing over $f(x)^\top \widehat{\theta}_t$. This set should include potential high reward arms in feasible space Ω_{t-1}, and thus samples from potential high reward regions in the entire sequence space of DAISY designs. To further explore these regions, *in silico* mutagenesis is performed over this hypothesis set to obtain a larger set of designs. This larger set of designs is then thresholded by the entropy prediction model, with the remaining designs added to Ω_{t-1} to obtain Ω_t. In implementation, we picked the top batch of designs from Ω_t that maximize the objective function above at each round (Figure 3G).

For experimental testing of our DAISY design predictions, we further filtered out designs that contain G quadruplexes and poly-T stretches that could introduce problematic secondary structure and inhibit PolIII-based transcription of the spacers targeting the DAISY barcode, respectively.

Feature extraction to identify motifs and rules defining high entropy DAISY barcodes

To understand how different input features contribute to entropy and identify motifs indicative of high entropy, we borrowed ideas from deep visualization tools in computer vision (Mahendran and Vedaldi, 2015). Specifically, we extracted 100 CLOVER sequences with the highest predicted entropy and calculated the gradients on the input features of the entropy prediction model. These gradient values are then averaged using the mean as the definition of center to rank the 4906 features (Table S3).

scDAISY-seq barcode sequencing analysis

Resulting fastq data were processed using CellRanger V4 (10x Genomics) using a custom feature barcode reference. The resulting.bam file was filtered to include only reads mapping to the custom feature barcode reference. Next, reads were collapsed into groups defined by their 10x Cell Barcode and UMI sequences. The collapsed reads were then parsed to extract only the evolvable barcode sequence and aligned to the reference evolvable barcode sequence using the Smith-Waterman algorithm with the following parameters: gapopen=13, gapextend=0.5. The alignment.bam file was then parsed to group the Cas12a-edited barcode alignment to a 10x Cell Barcode and UMI sequence – corresponding, in theory, to a unique transcript within a single cell. Cells were then clonally grouped into lineage groups based upon their static barcode sequence using hierarchical clustering. The phylogeny of cells within each lineage group was reconstructed using the Neighbor Joining algorithm (Jones et al., 2020).

Cell cycle scoring analysis

To quantify the contribution of cell cycle phases to the transcriptional clustering of cells, we scored each cell by the expression level of genes associated with either S phase or G2M phase using scanpy (Wolf et al., 2018). Briefly, the average expression (mean) of the S phase and G2M gene sets were computed for each cell and used to classify cells into S phase, G2M phase or G1 phase. The cell cycle scores were then overlayed onto the transcriptional clusters to identify clusters enriched for cells in a specific cell cycle phase.

Transcriptional memory analysis

Single-cell transcriptome profiles were generated from libraries sequenced with paired-end 150 cycle configuration on a HiSeq 4000 (Illumina) or NovaSeq 6000 (Illumina). Resulting fastq data were processed using CellRanger V4 (10x Genomics) using a custom feature barcode reference. Gene expression data were processed using scanpy (Wolf et al., 2018). Briefly, barcoded cells were selected from the raw gene expression matrix. Cells were filtered based upon their UMI abundance (100 UMI cutoff) and expression frequency (expressed in at least three cells). The data were further filtered to exclude cells in which mitochondrial genes represented greater than or equal to 20% of the total UMIs within the cell. Finally, the UMI counts were normalized on a per cell basis with the target sum set to 1×10^4 (transcripts / 1×10^4 molecules). Finally, the normalized counts were pseudo-counted by 1 and logarithmically transformed. The phylogenetic structure of each lineage group was used to assemble a dictionary in which all leaves within the tree were uniquely grouped by their most recent common ancestor (MRCA) into a sublineage (*s*). For each gene (*g*), a memory index (*m*) was calculated as follows:

Let μ_s be the mean expression of *g* within *s*

Let σ_s be the standard deviation of the expression of *g* within *s*, where

$$CV_s = \frac{\sigma_s}{\mu_s}$$

Across all sublineages (*S*), min(CV_s) was determined. To generate the null distribution, 1000 random phylogenetic trees were simulated with the same sublineage sizes as in the C1 tree (Figure 7F). The minimum *CV* within a random sublineage min(CV_{random}) was computed as described above. The mean min(CV_{random}) was determined across all simulations. The final memory index was defined as:

$$m = \text{mean}(\min(CV_{random})) - \min(CV_s)$$

To assess how the memory index changes in accordance with phylogenetic distance, cells were grouped together if they were separated by a node distance of 2 (cousins). The resulting cell groupings were used to calculate the memory index as previously described for the analysis using sister cell groupings.

Memory genes across three percentile thresholds (85th, 90th, and 95th) were selected for functional follow-up with gene set enrichment analysis (Table S6). Briefly, we utilized a publicly available GSEA software portal (Mootha et al., 2003; Subramanian et al., 2005) in which we queried each memory gene set for significantly enriched Gene Ontology (GO) biological components. Second, we utilized the enrichR package (Chen et al., 2013) to identify proteins with enriched ChIP-Seq peaks from publicly available ENCODE data. To follow-up we downloaded publicly available ChIP-seq data from A375 cells (GSE133834) (Su et al., 2019). Genes with a memory index greater than or equal to the 90th percentile (n = 1001 genes) were selected for visualization of EZH2 binding profiles using deeptools2 (Ramírez et al., 2016). To generate the TPM-matched gene set, bulk RNA-expression levels were determined for each memory gene in A375 cells using the CCLE (Barretina et al., 2012). Genes whose values were within +/- 20% of memory gene expression were included for visualization. Briefly, BED files containing the transcription start sites (TSSs) were generated for all memory and control gene sets and used to compute a matrix in which genomic regions are scored for enrichment in EZH2 binding. The intensity of EZH2 binding was visualized using the plotheatmap function within deeptools2.

TCGA analysis of EZH2 expression and transcriptional heterogeneity

RNA-Seq profiles of skin cutaneous melanoma (SKCM) lesions were downloaded from The Cancer Genome Atlas (TCGA). The Shannon entropy was then calculated using the distribution of expression levels of each gene expressed within a tumor sample to generate a transcriptional heterogeneity metric for each tumor. The *EZH2* expression level within each tumor was determined and the set of values was binned into four groups, within the range of expression values, for visualization of the relationship between the expression

level of *EZH2* in a tumor and its transcriptional heterogeneity. To benchmark the relationship between *EZH2* expression level and transcriptional heterogeneity, we calculated the Spearman rank correlation coefficient (SCC) of the linear relationship between the expression level of each gene within a tumor and the transcriptional heterogeneity of the tumor. Genes were ranked by the SCC and plotted.

Bioinformatic analysis of editing outcome of DAISY-chain barcodes

The DAISY-chain barcode sequencing data were analyzed using a custom pipeline. Briefly, reads were split into reads 1 and 2 and then trimmed and merged using *flash* v1.2.11 with default parameters (Martin, 2011; Magoč and Salzberg, 2011). The processed reads were then aligned to the reference DAISY-chain sequence using *needleall* (Rice et al.). The resulting alignments were then analyzed using custom python scripts. Briefly, these scripts quantified the representation of each alignment to call on-target edits that generate alleles used to estimate DAISY-chain barcode entropy using the Shannon entropy metric, as previously described. To estimate the Shannon entropy of the DAISY-chain barcode, the static tag was used to identify alleles that were present in only a single clonal population to remove noise.

Sequence evolution analysis of DAISY-chain barcodes

On-target edits were used to generate a dictionary of alleles assigned to clonal populations represented across each timepoint. A sequencing depth cutoff of 10 reads was used to remove noise generated due to sequencing error. The number of clones assigned to each allele was then calculated and plotted for each timepoint.