Using CRISPR barcoding as a molecular clock to capture dynamic processes at single-cell resolution

Yolanda Andres-Lopez[1], Alice Santambrogio[2,3], Ioannis Kafetzopoulos[2,3], Christopher Todd[2,3], Celia Alda-Catalinas[2,4], Stephen J. Clark[2,3], Wolf Reik[2,3] & Irene Hernando-Herraez[2,1]

[1] Department of Cells and Tissues, Instituto de Biología Molecular de Barcelona, Consejo Superior de Investigaciones Científicas (IBMB-CSIC), Barcelona, Spain

[2] Epigenetics Programme, Babraham Institute, Cambridge, United Kingdom

[3] Altos Labs, Granta Park, Cambridge, United Kingdom

[4] Genomic Sciences, GSK, Stevenage, United Kingdom

Correspondence to wreik@altoslabs.com and ihhbmc@ibmb.csic.es

Abstract

Biological processes are fundamentally dynamic, yet existing methods for capturing these temporal changes are limited. We present scDynaBar, a novel approach that integrates CRISPR-Cas9 dynamic barcoding with single-cell sequencing to enable the recording of temporal cellular events. In this system, genetic barcodes accumulate mutations over a 4-week timeframe and then are sequenced together with the transcriptome of each single cell. We propose that this gradual accumulation of genetic diversity can be exploited to create a time-ordered record of cellular events. To

demonstrate this, we apply the system to track the transition from a pluripotent state to a two-cell (2C)–like state in mouse embryonic stem cells (mESCs). The results provide compelling evidence for the transient nature of the 2C-like state. Additionally, our system shows consistent mutation rates across diverse cell types in a mouse gastruloid model, underscoring its robustness and versatility across various biological contexts. This technique not only improves our ability to study cellular dynamics but also creates exciting new opportunities for future applications based on recording temporal signals at the single-cell level—in other words, using dynamic barcoding as a molecular clock.

Introduction

Cellular behaviour is intrinsically dynamic, undergoing alterations in response to internal and external stimuli. Different methodologies are commonly used to track or infer these changes. However, these methods present several limitations. For example, methods based on live-cell fluorescence microscopy are often incompatible with *in vivo* studies and are restricted in terms of long-term and throughput capabilities[1]. Other methods, such as those based on inference algorithms using single-cell RNA sequencing (scRNA-seq) have low temporal resolution, as scRNA-seq only captures a snapshot of gene expression at one specific time point[2,3].

DNA memory systems based on CRISPR technology have emerged as valuable tools for recording dynamic events[4–6]. These systems exploit CRISPR-based techniques to induce permanent genetic modifications at precise genomic locations. By regulating the delivery of CRISPR components temporally, these systems create DNA mutations that act as molecular barcodes, which enables researchers to capture cellular activities retrospectively. Various DNA memory devices have been developed to track a wide range of biological signals[7], including gene expression[8], chemical exposure[9,10], lineage of origin[11–13], clonality[14,15], activation of signalling pathways[16] and inflammatory responses[17,18].

Despite significant advancements, single-cell resolution still must be achieved to gain a comprehensive understanding of cellular behaviour. DNA barcoding systems employing guide RNAs (gRNAs) that target polyadenylation (polyA) sites have enabled simultaneous sequencing of single-cell transcriptomes alongside DNA barcodes[19,20]. This integrated approach has been successfully applied to capture lineage of origin in developmental[21–27] and cancer studies[28–30]. However, single-cell systems are often limited by a temporal activity window of 24 h to 48 h[22,26], which

greatly restricts their capacity to record long-term cellular processes. We propose that extending this temporal window would significantly expand the technology's applications and enable it to function as a molecular recorder for a broader array of cellular signals and processes.

In this study, we present scDynaBar, a robust method for long-term barcoding with highly efficient single-cell capture. We demonstrate the versatility of scDynaBar by successfully tracking the transition from 2C-like states to pluripotency in mESCs in vitro. This approach revealed that the 2C-like state is transient and does not persist over extended periods. Additionally, our research shows consistent mutation rates across various cell types in mouse gastruloids, highlighting the broad potential of our approach for diverse biological contexts at single-cell resolution.

Results

Dynamic barcoding

We aimed to develop a versatile system that can (1) record molecular events over extended periods, and (2) simultaneously profile genetic barcodes and gene expression at the single-cell level. To achieve this, we engineered self-targeting gRNAs, which are a variant of the canonical form of CRISPR-Cas9 gRNAs with the unique ability to repeatedly target their own sequence, resulting in continuous mutagenesis over time[31]. In order to recover these genetic mutations, we engineered the system to retrieve them from a polyadenylated transcript (expressed barcode). We also included an mCherry reporter driven by a constitutive CAG promoter, which enables easy visualization and selection of targeted cells (**Fig. 1a**).

To investigate the dynamics of the system, and to test its ability to detect the barcodes at the RNA level, we performed a time-course experiment using seven different gRNAs with varying protospacer sizes (21 or 26 base pairs [bp]) and nucleotide compositions (**Supplementary Fig. 1**). We also aimed to evaluate the relative efficacies of barcoding using standard Cas9 and a base editing system, as the latter avoids inducing double-strand breaks. Specifically, we used the BE3 base editor, which comprises the Cas9 nickase fused to the cytidine deaminase APOBEC and the uracil DNA glycosylase inhibitor (UGI) domain[32]. This system facilitates the conversion of C:G base pairs to T:A base pairs in a 4 to 8 nucleotide region in the absence of DNA double-strand breaks. In our approach, the UGI domain of the BE3 system was removed to enhance the diversity of mutations (APOBEC–XTEN–dCas9(A840H))[32]. Both Cas9 constructs contained a constitutive promoter (CMV early enhancer fused to modified chicken β-actin promoter) and a GFP reporter. Expression of the constructs was controlled by a Cre-responsive FLEX switch, ensuring that expression occurred only after recombination. We co-transfected individual gRNAs together with one of the two Cas9 constructs into mESCs. After stable integration, we activated Cas9 by adding Cre recombinase to the culture media, and then collected cells at various time points, up to day 18 (**Fig. 1b**). We then selected cells positive for both reporters (GFP and mCherry) by fluorescence-activated cell sorting (FACS) (and subsequently performed bulk amplicon sequencing on their RNA (see **Supplementary Fig. 2 and Methods**) (**Fig. 1b**).

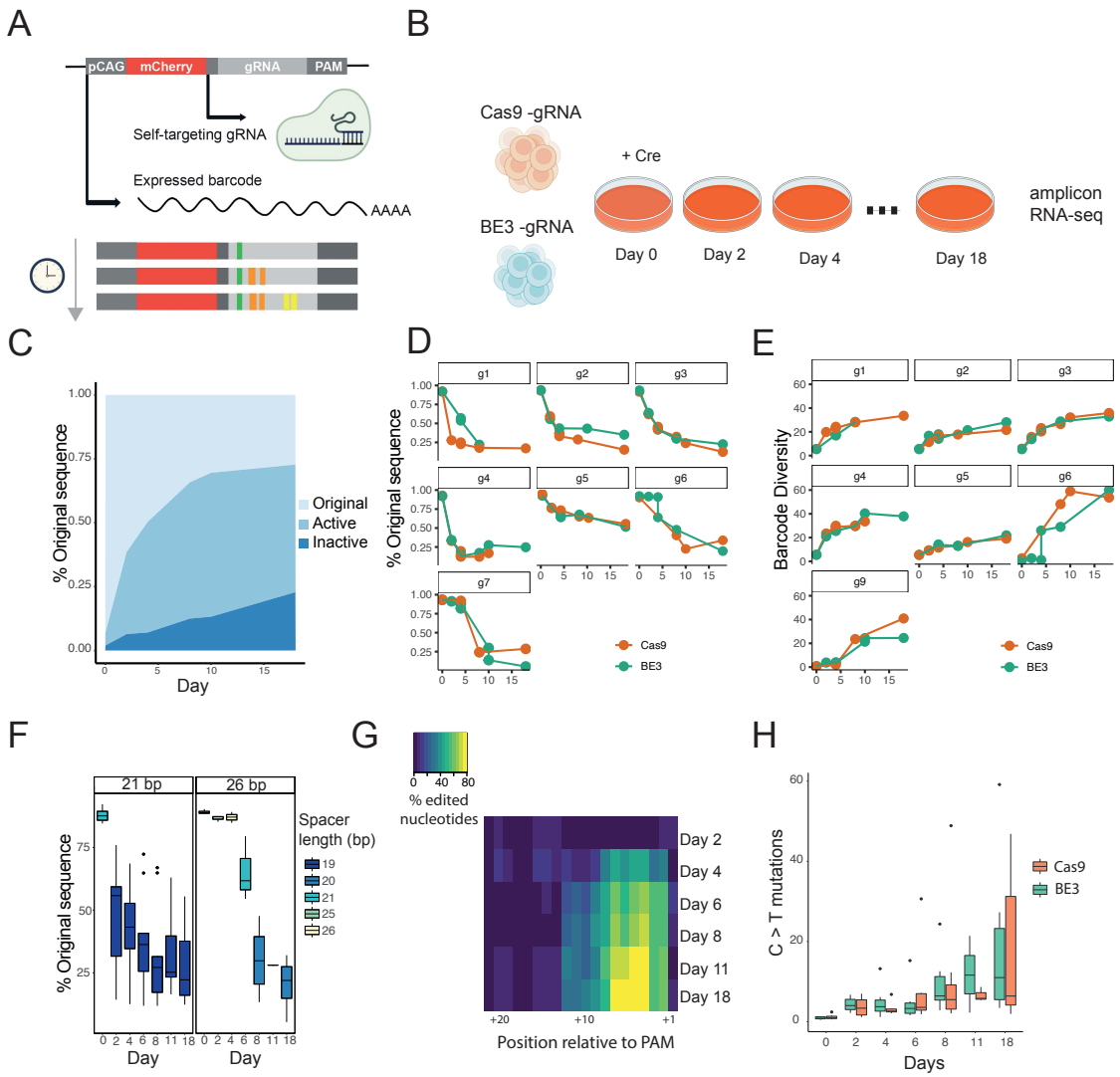

Figure 1. a. *Schematic representation of dynamic barcoding with a poly-A readout.* ***b.*** *Schematic representation of the experimental setup.* ***c.*** *Proportion of barcodes based on their mutational profile, combining data for both systems: edits on the protospacer with intact PAM motif (active); absence of PAM motif (inactive); and uncut gRNA (original).* ***d.*** *Proportion of original barcodes over time across different gRNAs.* ***e.*** *Barcode diversity over time, considering mismatches, gaps and gap extensions across different gRNAs.* ***f.*** *Proportion of original barcodes over time for gRNAs with 21-bp spacers (left) or 26-bp spacers (right). Box plots are coloured by the mean spacer length at different time points (Cas9 system).* ***g.*** *Percentage of the original nucleotide over time, aligning the spacer relative to the PAM sequence (Cas9 system).* ***h.*** *Percentage of C>T mutations over time, considering all different gRNAs classified by Cas9 versions. For all boxplots, the boxes represent the interquartile range (IQR), with the horizontal line inside each box indicating the median.*

To investigate the general dynamics of the systems, we first classified the barcode sequences into three groups: (1) the original barcode, in which no editing occurred, (2) an active barcode, in which editing occurred in the protospacer region but with the PAM sequence still uncut, and (3) an inactive barcode, in which editing occurred in both the protospacer regions and the PAM sequence, preventing further mutagenesis.

Our data are consistent with sequential editing, as they show increased level of active and inactive gRNAs, and a decreased frequency of original barcodes, over time (**Fig. 1c**). The increased editing rate is consistent when using any of the seven gRNAs that we tested, as well as with either version of Cas9 (**Fig. 1d**) (Jonckheere-Terpstra test; $p = 2 \times 10^{-13}$). Background editing in the absence of Cre (at day 0) was low (averaging 3% across experiments). After 18 days of editing, the system still contained barcodes with no mutations (mean original barcode frequency, 26.8%), indicating that mutagenesis could potentially be extended beyond this day (**Fig. 1d**). We next assessed the diversity of the edited reads by comparing their sequences to the original barcode. In this analysis, we performed pairwise alignments for over 100,000 barcode sequences and calculated a diversity score considering mismatches, gaps and gap extensions. All seven gRNAs demonstrated a sequential increase in diversity over the 18 days (Jonckheere-Terpstra test; $p = 8 \times 10^{-16}$), providing strong evidence for continuous mutagenesis as well as for their capacity to effectively record molecular events (**Fig. 1e**). In agreement with previous studies, we also observed that in the Cas9 system 26-bp gRNAs exhibit a slower editing rate compared to 21 -bp gRNA during the initial four days (**Fig. 1f**)[31] (Day 4, Mann-Whitney-Wilcoxon test; $p = 0.0005$). After this time point, the frequency of intact barcodes for both the 21-bp and the 26-bp barcodes became very similar. This could be attributed to the accumulation of deletions in the 26-bp gRNAs (**Fig. 1f**), potentially leading to an increased editing rate at later time points.

Next, we analysed the distribution of mutagenesis relative to the PAM of each gRNA in the Cas9 system. As expected from the activity of Cas9, we observed a preferential accumulation of mutations near the cut site, specifically within three nucleotides upstream of the PAM site (**Fig. 1g**). We then compared the specific types of edits in

the BE3 system to the standard Cas9. Notably, we observed a similar accumulation of indels over time in both systems (Mann-Whitney-Wilcoxon test; $p = 0.3$, **Supplementary Fig. 3**). This could be attributed to the absence of the UGI domain, which may lead to DNA repair by alternative cellular mechanisms, ultimately resulting in indel formation. Despite the lack of the UGI domain, the BE3 system exhibited a slightly higher frequency of C>T mutations at most time points, with an average difference of 6.9% (**Fig. 1h**).

Reproducibility and window of activity

We next aimed to extend the duration of mutagenesis to determine the maximum period during which the system can remain active and continue to generate mutations. As before, each of the seven gRNAs was individually co-transfected with one of the Cas9 constructs and then processed at various time points, up to day 32 (**Fig. 2a**). Remarkably, we observed consistent results between biological replicates for the initial 18 days, with a Pearson's correlation coefficient of 0.94 ($p < 2.2 \times 10^{-16}$), indicating the high reproducibility of our system (**Fig. 2b**). These results also indicate that the Cas9 system remained active until day 32, albeit working at a slower rate (**Fig. 2c, d**). In contrast, the BE3 system exhibited a significant increase in the frequency of original barcodes and a decrease in sequence diversity at day 25 (**Fig. 2c, d**). This could be caused by a bottleneck effect, where non-edited cells outcompeted edited cells due to potential deleterious effects[33]; in any case, this makes the Cas9 system more suitable for long-term applications. We also observed that a proportion of gRNA insertions appeared to resist editing, as indicated by the presence of original barcode sequences in both systems at day 32 (**Fig. 2c)**.

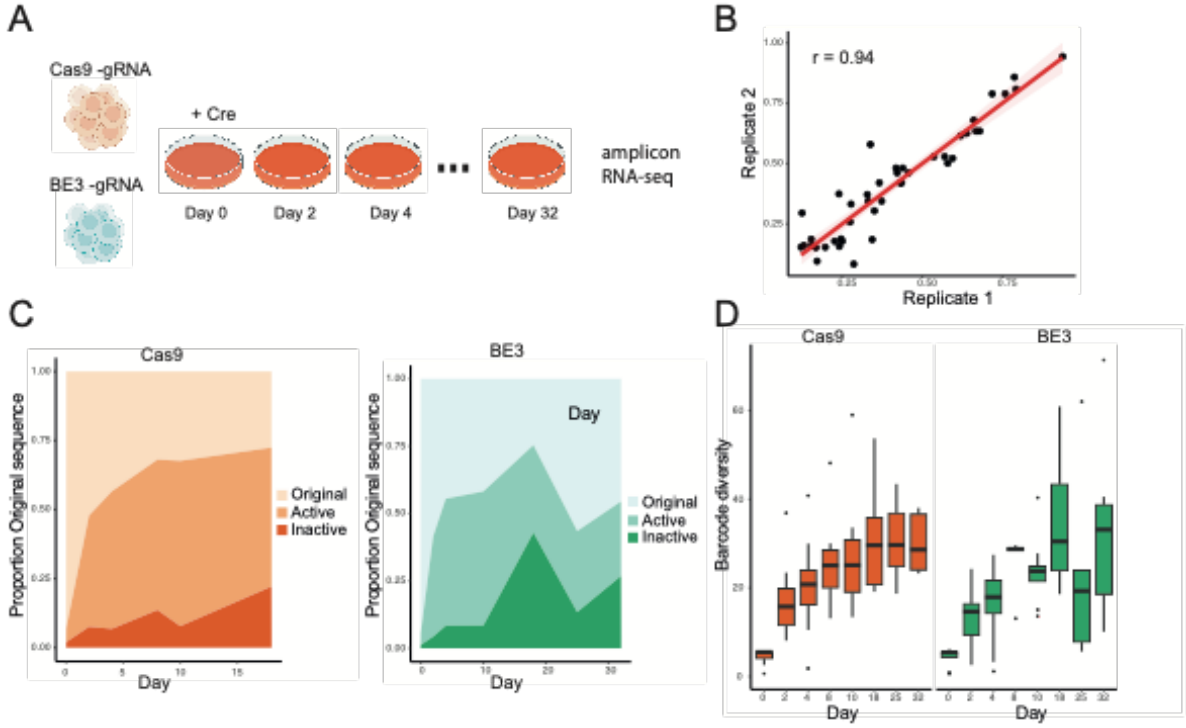

*Figure 2. **a**. Schematic representation of the experimental setup. **b**. Pearson's correlation coefficient for reproducibility between replicates from days 0 to 18. **c**. Proportion of barcodes based on their mutational profile: edits on the protospacer with intact PAM motif (active); lack of PAM motif (inactive); and uncut gRNA (original) according to Cas9 (left) or BE3 (right). **d**. Barcode diversity over time, considering mismatches, gaps and gap extensions, pooling all gRNA for Cas9 (left) and BE3 (right). In the boxplots, the boxes represent the IQR with the horizontal line inside each box indicating the median.*

Simultaneous detection of genetic barcodes in single cells

We then evaluated whether genetic barcodes could be simultaneously recovered together with transcriptomic data using standard single-cell RNA sequencing. First, based on our previous observations, we used gRNA g3 together with the canonical form of Cas9 to establish a monoclonal cell line. We grew the cells under serum/LIF conditions; following Cre induction, as we FACS-sorted double-positive cells (for GFP and mCherry) on days 4 and 10 (**Fig. 3a**). Additionally, we included a non-induced group that showed minimal Cas9 expression at day 10 (GFP = 0.1%), which confirmed the absence of background activity.

Cells were processed using the 3' 10X Genomics platform, where they were encapsulated in droplets with barcoded PolyT-coated beads and subjected to whole transcriptome amplification following the standard 10x Genomics scRNA-seq 3′ protocol. We then specifically amplified and sequenced the genetic barcodes from the cell-barcoded cDNA (**Fig. 3a**). Overall, we successfully recovered genetic barcodes from 97% of cells that passed transcriptome quality control, demonstrating the efficiency of our approach for simultaneous transcriptome and barcode detection.

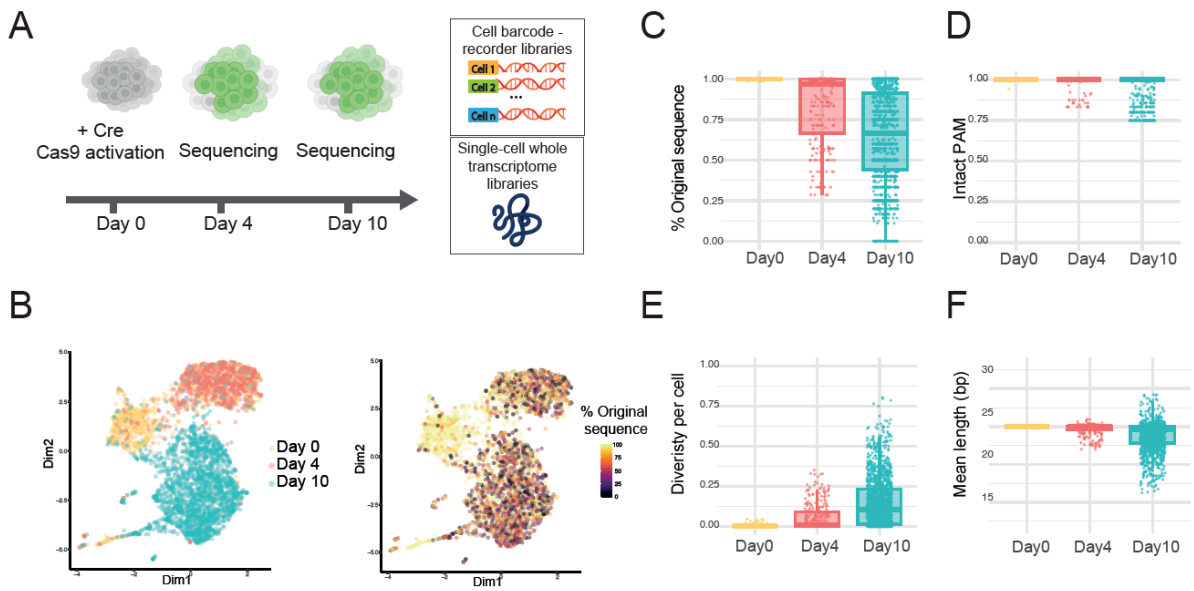

Figure 3 a. *Schematic representation of the experimental setup.* ***b****. t-SNE plot coloured by cluster: non-induced (day 0) and induced (day 4 or day 10) (left). Colour indicates the mean percentage of the original barcode sequence per cell (right).* ***c–f****. Boxplot of results at days 0, 4 and 10 showing* ***c)*** *the percentage of the original barcode sequences per cell;* ***d****) the mean proportion of barcodes with full PAM sequence per cell;* ***e****) mean length of the barcodes per cell and* ***f****) mean sequence diversity per cell. For all boxplots, the boxes represent the IQR, with the horizontal line inside each box indicating the median. Each dot represents an individual cell, with data points displayed within the 95% confidence interval.*

t-SNE analysis revealed three clusters of cells, corresponding to the different time points: non-induced (day 0; *n* = 495), induced at day 4 (*n* = 626) and induced at day 10 (*n* = 1117) (**Fig. 3b**). The most frequent barcodes sequences are listed in (**Supplementary Table 1),** illustrating the progressive increase in edited barcodes over time. For each cell, we assessed the level of barcoding by analysing: the proportion of reads carrying the original barcode sequence, the mean barcode

diversity score (calculated as described previously), the average length of the barcode sequence, and the percentage of barcodes carrying a PAM sequence (**Fig. 3c–f**). Our findings align with the results in bulk, showing sequential editing over time (Jonckheere-Terpstra test; $p = 2 \times 10^{-04}$, 1000 permutations). We observed a decrease in the presence of original sequences along with a decline of the proportion of barcodes carrying a PAM motif (**Fig. 3c, d**). Additionally, we observed an increase in barcode diversity and greater variability in their lengths (**Fig. 3e, f).** We also observed that the top 100 most commonly edited gRNAs sequences were unlikely to have off-targets effects that could affect cell viability (**Supplementary Table 2**). Altogether, these data demonstrate the effectiveness of dynamic barcoding in single cells and the capability of the method to simultaneously capture barcodes and transcriptomic information using standard single-cell technologies.

Recording dynamic transitions

Having established the efficacy of scDynaBar in single cells, we next applied it to track transitions between pluripotency and totipotency-like states. Under standard serum/LIF culture conditions, a small population (less than 1%) of mESCs spontaneously emerges that exhibits features resembling the totipotent 2C stage[34,35]. These features include a specific transcriptional profile, such as the expression of Zscan4 genes, increased histone mobility, DNA demethylation and the capacity to contribute to both embryonic and extraembryonic tissues[34,36,37]. Although it is believed that the 2C-like state cannot self-propagate but rather continuously undergoes transitions to a pluripotent state, these dynamics have not been directly measured[34,38]. To track 2C-like conversions, we integrated our barcoding approach with a pZscan4-CreERT2 construct (**Fig. 4a**)[39]. With this setup, cells entering the 2C-like state

permanently activate the barcoding cassette during treatment with tamoxifen. After establishing a monoclonal line, cells were treated with tamoxifen for 12 days. Following treatment, GFP- and mCherry-positive cells were FACS-sorted (GFP- and mCherry-positive) and processed for single-cell RNA-seq and barcoding sequencing (**Fig. 4a**). We also sequenced a portion of GFP-negative cells as a control.

After cell and barcode quality control, we recovered 3,352 GFP-positive cells and 1,270 Cas9-GFP-negative cells (barcode recovery rate = 98%). t-SNE analysis identified three clusters corresponding to Cas9-GFP-negative cells and two clusters representing Cas9-GFP-positive cells. Within the Cas9-GFP-positive population, we observed variable editing rates, indicating the activation of barcoding at distinct time points during our time course (**Fig. 4b**). t-SNE analysis also revealed a cluster of 24 cells corresponding to 2C-like cells, characterized by expression of totipotency markers, such as Zscan4c (**Fig. 4b and 4c).**

Cells in the pluripotent state, characterized by a lack of Zscan4c expression, exhibited high variability in their barcode diversity (**Fig. 4b**). This high variability indicates that the CRISPR barcoding system was activated early upon entry into the 2C-like state, at different time points in each of the cells. As a result, the continuous activity of the barcoding system led to an accumulation of different rates of diversity over time. In contrast, all cells in the 2C-like state (i.e., with Zscan4c expression) displayed low diversity in their barcodes, along with a high proportion of the original barcode sequence (**Fig. 4d**). The absence of 2C-like cells with high barcode diversity indicates that they transitioned into this state only recently. These findings prove the prevailing hypothesis in the field that ESCs undergo continuous dynamic transitions into and out of the 2C-like state, a phenomenon that has been difficult to measure directly until

now[34,38,40]. Our results also indicate that the 2C-like cells that became pluripotent did not re-enter into the 2C-like state during the experiment. This could be attributed either to probability, as only about 1% of the cells are in the 2C-like state at any given time, or to potential inhibitory signals that might delay re-entry into this state.

These results demonstrate the effectiveness of scDynaBar in capturing the dynamic transitions between these states, making it a versatile tool for studying complex cellular dynamics that could not be previously tracked.

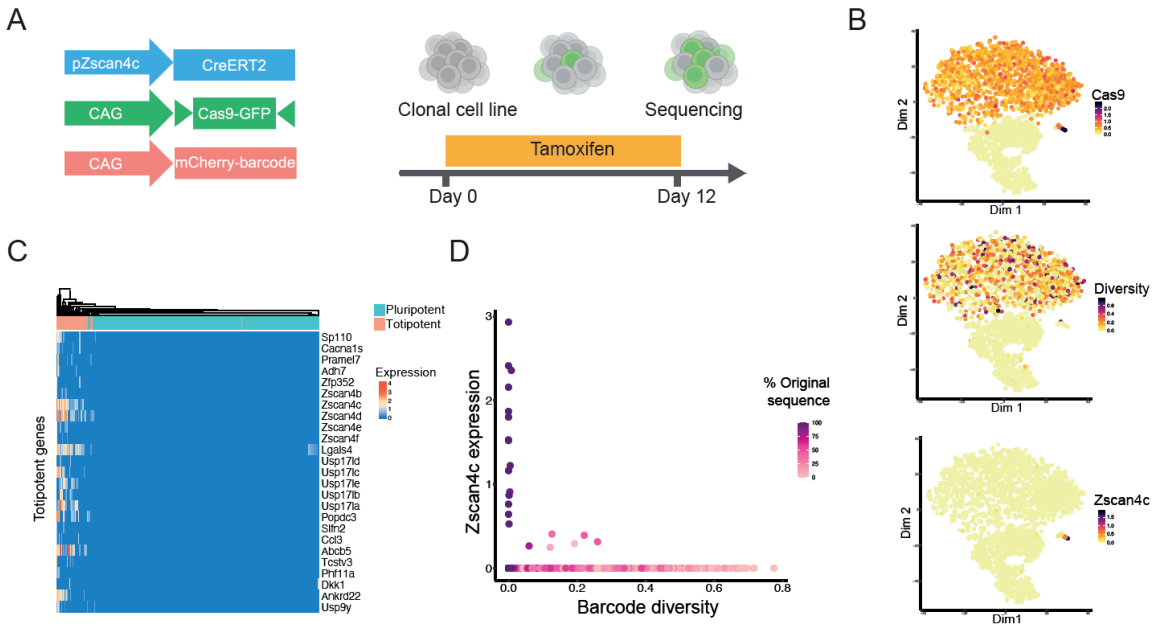

Figure 4. ***a.*** *Schematic representation of the experimental setup for tracking 2C-like state transitions.* ***b.*** *t-SNE plot coloured by (top) Cas9 expression, (middle) mean barcode diversity per cell, and (bottom) Zscan4c expression.* ***c*** *Heatmap displaying gene expression for top markers for the totipotent cluster.* ***d.*** *Scatterplot displaying Zscan4c expression (y-axis) against barcode diversity (x-axis) per cell, with colour indicating the percentage of original barcode sequences.*

Dynamic barcoding in mouse gastruloids

To investigate the mutation rate across diverse cell types, we applied scDynaBar to mouse gastruloids, which are three-dimensional structures that mimic early developmental processes and comprise a diverse array of cell types[41]. For this, we used our previously established mESC monoclonal cell line (**Fig. 3**). Two days after Cre induction, mESCs were seeded for aggregation and subsequent gastruloid

formation[41] (**Fig. 5a**). Six days after induction, Day 4 gastruloids were dissociated, and the cells were processed for single-cell RNA and barcode sequencing. After the cell and barcode quality controls, we recovered 1,057 GFP-positive cells (i.e., GFP counts >1) and assigned cell types by mapping RNA expression profiles to a reference atlas of mouse embryos, from E6.5 to E8.5[42] (**Fig. 5b**). This analysis revealed a bias towards the mesodermal lineage, consistent with our previous studies showing that gastruloids often favour mesodermal or ectodermal cell types[43]. We then computed the proportion of the original barcode sequence for each cell and found no clear differences between cell populations (**Fig. 5c**), indicating that the barcoding rate is similar across diverse cell types. Comparing these data with our previous dataset (**Fig. 3**), we observed that gastruloids collected at day 6 exhibited levels of barcoding that ranged between those observed at day 4 and day 10 (**Fig. 5d**). Specifically, the median value of the original barcode sequencing was 80% at day 6, compared to 75% at day 4 and 60% at day 10, indicating an accumulative increase of mutations over time (Jonckheere-Terpstra test; $p = 2 \times 10^{-04}$; 1000 permutations) (**Fig. 5d**).We hypothesized that the gradual increase in barcode editing could be used to determine the duration of barcode activity in single cells. To test this hypothesis, we developed a temporal predictor using random forest regression on our dataset (days 0, 4, 6 and 10). The predictor was trained on 80% of the data, incorporating barcode features as variables (mean barcode length, percentage of original sequences, diversity mean and coverage). Using this predictor, we achieved an R^2 of 98% on the training set, and of 95% on the test data (**Fig. 5e**). Achieving this high level of accuracy despite using a limited number of time points suggests that DNA barcodes can effectively be used to predict the duration of barcoding in single cells. This highlights the strong potential of

dynamic barcoding to function as a molecular clock for precise time tracking in different cellular contexts.

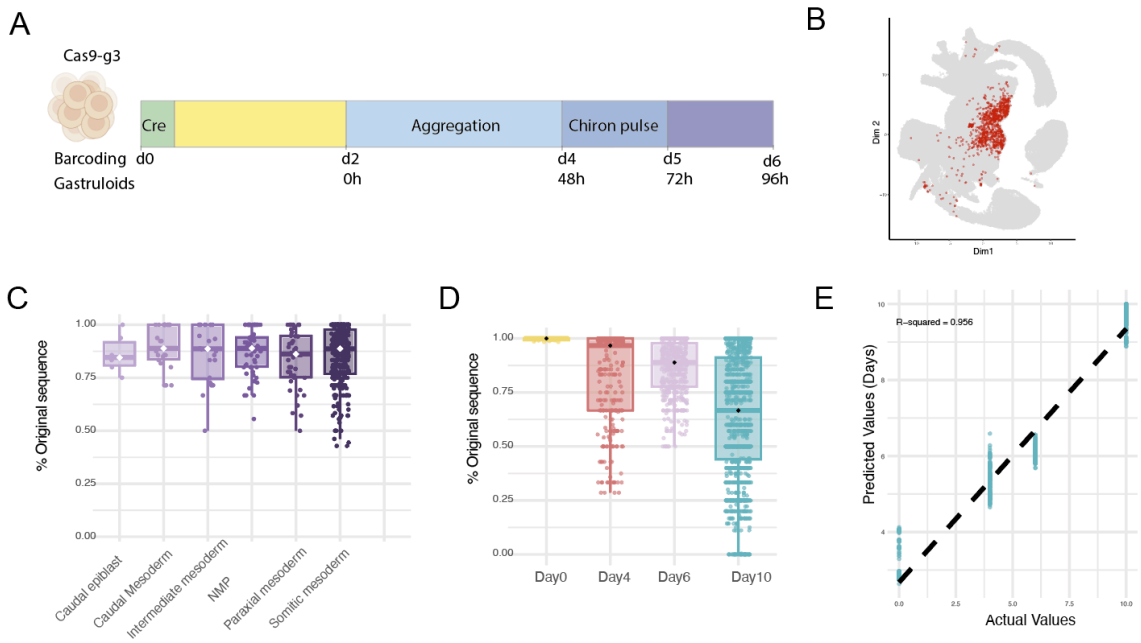

***Figure 5. a**. Schematic of the experimental design for gastruloid formation and barcoding. **b**. UMAP projection showing developmental stages from a reference atlas of mouse embryos between E6.5 and E8.5, with extra-embryonic cells excluded. Gastruloid cells are highlighted in red. Nearest neighbour analysis was used to assign cell-type labels to the gastruloid cells based on their proximity to the cells in the reference atlas. **c**. Boxplot showing the percentage of the original barcode sequence was calculated for each cell grouped by cell type. **d**. Boxplot showing the distribution of the proportion of original barcodes across different time points. **e**. Scatter plot showing predicted (y) versus actual (x) time points (days) using a random forest model, with R-squared values of 95% for test data. For all boxplots, the boxes represent the IQR with the horizontal line inside each box indicating the median.*

Discussion

In this study, we introduce scDynaBar, a novel approach for dynamic cellular barcoding in single cells. We achieved this by engineering self-targeting gRNAs[31], enabling the simultaneous capture of both genetic barcodes and the transcriptomes of individual cells. This method requires minimal adjustments to existing single-cell sequencing workflows offering a high level of both versatility and ease of scalability.

We demonstrated that dynamic barcoding leads to the sequential accumulation of genetic diversity over four weeks through continuous targeting, with high

reproducibility. This represents a substantial advancement as compared to other single-cell barcoding methods, in which mutations typically occur within the first 48 h[22,26]. The accumulation of mutations over time can be assessed through multiple measures, including i) a reduction in the proportion of original sequences, ii) a decrease in the frequency of sequences with an intact PAM motif and iii) a notable increase in both the diversity of spacer sequences and iv) the variability of their lengths. We observed comparable rates of mutagenesis using standard Cas9 or the cytosine base editor BE3[32], which fuses APOBEC with nCas9; however, we noted potential adverse effects associated with BE3 during prolonged barcoding.

As a proof of concept for scDynaBar's potential applications, we demonstrated its ability to record dynamic cell transitions. By collecting data at a single endpoint and using a CreERT2 protein controlled by the Zscan4c promoter, we documented the transition of mESCs from the pluripotent state to the 2C-like state. In this experimental setup, the diversity of barcodes encodes the timing of the cell transitions to the 2C-like state. Our findings showed that all 2C-like cells had recently transitioned to this state, as evidenced by a lack of mutations in their barcodes. These results underscore the transient nature of the 2C-like state and support the current hypothesis suggesting that the 2C-like state cannot self-perpetuate[34,38,40]. Additionally, these findings illustrate the potential of our approach for studying complex cellular dynamics that are currently difficult to track with existing methodologies.

One exciting aspect of scDynaBar is its remarkable adaptability to different experimental conditions. Our observations from a mouse gastruloid model revealed a consistent mutation rate across various cell types, demonstrating the robustness of our approach and suggesting that scDynaBar can be reliably applied to diverse

biological contexts. We believe this approach can be extensible, making it suitable to be used to record a wide variety of temporal signals and processes in single cells. For example, by modulating Cas9 activity, scDynaBar could be tailored to track specific transcriptional events, such as responses to external stimuli or complex disease phenotypes. This versatility paves the way for new research avenues in capturing cellular signals and dynamic processes at the single-cell level.

Methods

Vector construction

The barcoding cassette used in this study were constructed by incorporating gBlock (IDT DNA) synthesized DNA fragments, including CAG promoter, mCherry, and scaffold, into a piggyBac backbone. Different gRNAs (spacers), obtained from Sigma-Aldrich, were cloned into this vector. The sequences of these spacers are provided in the **Supplementary Fig. 1**. The Cas9-2A-GFP vector used in this study is based on Cong et al. (2013)[44] and was cloned into a piggyBac backbone with a one-way genetic switch (FLEX) system. The BE3 system, kindly provided by Alexis Komor[32], was further modified by removing the uracil glycosylase inhibitor (UGI) and then incorporating it into a piggyBac backbone as nCas9-rAPOBEC1-2A-GFP in a one-way genetic switch (FLEX) system.

The plasmid carrying CreERT2 under the control of the Zscan4c promoter (pZscan4-CreERT2 cells), as described in Zalzman et al. (2010)[39], was generated by VectorBuilder with a puromycin resistance marker.

Mouse embryonic stem cell culture

E14 mESCs were cultured under standard serum/LIF conditions in DMEM (Gibco, 11995-040) supplemented with 15% fetal bovine serum, 1 U/ml penicillin and 1 mg/ml streptomycin (Gibco, 15140-122), 0.1 mM nonessential amino acids (Gibco, 11140-050), 4 mM GlutaMAX (Gibco, 35050-061), 50 μM β-mercaptoethanol (Gibco, 31350-010) and 103 U/ml LIF (Stem Cell Institute, Cambridge). Cells were maintained at 37°C in a 5% CO_2 atmosphere on gelatinized tissue-culture plates. Media was changed daily, and cells were passaged every other day using trypsin-EDTA (Thermo Fisher Scientific, 25200056).

Stable cell lines were generated through transfection using Fugene6 (Promega, E2311), followed by drug selection and fluorescence-activated cell (FAC) sorting. Clonal cell lines were isolated through additional rounds of subcloning. Cas9 induction was carried out using Cre recombinase (Takara, 631449) in media supplemented with 6 μg/mL polybrene (Sigma, H9268), according to the manufacturer's instructions. pZscan4c-CreERT2 cells were treated with 4-hydroxytamoxifen (1 μM) for 12 days (Sigma, H7904).

Gastruloid culture

The mESC cell line containing the barcoding system was cultured under standard serum/LIF conditions. Induction was carried out as before using Cre recombinase (Takara, 631449) in media supplemented with 6 μg/mL polybrene (Sigma, H9268), according to the manufacturer's instructions. After two days, cells were prepared for gastruloid formation using a previously described protocol[41]. Specifically, mESCs were dissociated into single cells with trypsin-EDTA (Thermo Fisher Scientific, 25300096).

Cells were then washed twice with pre-warmed PBS (Thermo Fisher Scientific, 14190144), and the pelleted cells were then resuspended in 5 ml of N2B27 medium of a 1:1 mix of DMEM/F12 (Thermo Fisher Scientific, 11320033) and Neurobasal (Thermo Fisher Scientific, 21103049), supplemented with 0.5× N-2 Supplement (Cell Therapy Systems, A1370701), 0.5× B-27 Supplement (Thermo Fisher Scientific, 17504044), 2 mM GlutaMAX, 10 U/ml penicillin-streptomycin and 0.1 mM 2-mercaptoethanol. Cells were then diluted to a density of 7500 cells/ml in N2B27 medium, and 40 µl of this dilution was added to each well of a U-bottom 96-well suspension culture plate (Greiner Bio-One, 650185) to reach a density of 300 cells per well. The cells were left to aggregate for 48 h. After this period, 150 µl of N2B27 medium containing 3 µM CHIR99021 (Department of Biochemistry, University of Cambridge) was added. Every 24 h, 150 µl of the medium was replaced with fresh N2B27 medium without CHIR99021. On day 4, gastruloids were harvested for sequencing: they were transferred to an Eppendorf tube, rinsed with PBS, and dissociated into single cells using Accutase (StemPro, A1110501). Cells were then washed twice with 5 ml of PBS containing 0.04% BSA (Gibco, 15260037) to remove the Accutase, followed by filtration through a 50 µm strainer (Sysmex, 1050553). The cell count and viability were determined using the Countess II Automated Cell Counter prior to single-cell sequencing.

Bulk barcode sequencing

Following induction, cells were sorted at various time points using either the BD Aria III or the BD Influx High-Speed Cell Sorter. About 8K to 12K cells positive for GFP and mCherry were collected for each sample. RNA was then extracted using the RNeasy Micro Kit (Qiagen, 74004). To amplify the gRNA loci, reverse transcription was

performed as previously described[45], with 1 µl of primer (GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-T(30)) and 500 ng of RNA. cDNA was amplified in a 20 µl reaction using KAPA HiFi HotStart ReadyMix (KAPA Biosystems, KK2502) with the following primers:

Forward: 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCT(N/NN/NNN)TCTTGTGGAAAGGACGAAACAC-3′

Reverse: 5′-CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3′

XXXXXX- index;

Random nucleotides (NNN, NN and N) were added to introduce sequence diversity, enhancing variability over the constant sequence for sequencing purposes. The DNA product was then purified with a 0.8:1 volumetric ratio of AMPure XP Beads (Beckman Coulter, A63881) and the whole volume was resuspended in 20ul of PCR mastermix (KAPA HiFi Readymix) using the primers:

Forward: 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′

Reverse: 5′-CAAGCAGAAGACGGCATACGAGAT-3′

PCR products were purified using a 0.8:1 volumetric ratio of AMPure XP Beads, pooled, and sequenced on a single end Illumina MiSeq run with 58 cycles and an 8-cycle index.

Single-cell sequencing

A single-cell suspension was loaded into the 10X Chromium device, and libraries were prepared using the 10X Single-Cell 3′ Library & Gel Bead Kit v2 (10X Genomics, PN-120237), following the manufacturer's instructions. The following samples were each loaded into a lane of the 10X Chromium Control chip: day 4, day 10, pZscan4c-CreERT2 cells, gastruloids 1, gastruloids 2, and gastruloids 3.

Amplicon gRNA PCRs (KAPA HiFi Readymix) were performed for each of the 10X cDNA samples using 1 µl of the cDNA and the following primers:

Forward: 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCT(N/NN/NNN)TCTTGTGGAAAGGACGAAACAC-3′

Reverse: 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′

As before, random nucleotides (NNN, NN and N) were added to introduce sequence diversity, enhancing variability over the constant sequence for sequencing purposes. PCR products were purified using a 0.8:1 volumetric ratio of AMPure XP Beads, and the whole volume was loaded into a second nested PCR (KAPA HiFi Readymix) with the following primers:

Forward: 5′-
CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTGACTGGAGTTCAGACGTGT
GCTCTTCCGATCT-3′

Reverse: 5′-
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG
ATCT-3′

XXXXXX- index

After purification using AMPure XP beads, the enriched gRNA libraries were sequenced together with the transcriptome libraries on the Illumina NovaSeq platform.

Bulk barcode analysis

A total of 160 samples were processed for bulk barcoding (see **Supplementary Table 3**). After sample demultiplexing, reads that did not contain both the flanking regions on the U6 promoter and the spacer regions on either side of the cassette were discarded ("AACAC" and "TTAGAG" respectively). Barcodes with mean Illumina Phred scores below 28 were also removed. Additionally, samples with fewer than 200 reads were excluded. After this quality control, a total of 151 samples were processed for further analysis.

The percentage of original sequences was determined by comparing the spacer regions to the original spacer sequences. To assess diversity, each read was subjected to pairwise alignment with the reference spacer using the pairwiseAlignment function from the Biostrings R package[46]. The alignment was performed using a global-local approach with a substitution matrix that assigned a score of –1 for

mismatches and +2 for matches, with gap extension and opening penalties set to 5. Alignment scores were normalized across all conditions. Aligned sequences were further analysed to compute nucleotide frequencies per position.

Single-cell transcriptome quality control and processing

All 10X scRNA-seq data were processed using the 10X Genomics Cell Ranger pipeline with the mm10 genome build, including the GFP sequence. For quality control, cells were discarded if they had fewer than 3,000 detected genes and/or more than 7.5% of UMI reads originating from mitochondrial genes (**Supplementary Fig. 4**). Raw counts for each cell were normalized by dividing the counts by their size factors, and the resulting normalized counts were used for further analysis.

Broad clustering analysis and t-distributed stochastic neighbour embedding (t-SNE) plot generation were performed using Seurat. Linear dimensional reduction was executed via principal component analysis (PCA) using the Seurat function RunPCA, with highly variable genes identified by the Seurat function FindVariableGenes as inputs. Fifteen principal components were utilized to generate clusters and t-SNE plots using the Louvain algorithm implemented in the Seurat function FindClusters.

Gastruloid cell type assignment

Cell types were assigned by aligning RNA expression profiles to a reference atlas as previously described[43,47]. Briefly, count matrices from both datasets were combined and normalized. Highly variable genes were identified and used for PCA. Batch correction was performed to eliminate technical differences between the query and atlas cells. Using the integrated data, a k-nearest neighbours (kNN) graph was constructed. Each query cell's type was assigned by determining the most common

cell type among the 30 nearest neighbours in the atlas, using a Dirichlet distribution for majority voting.

Single-cell barcode library processing

Fastq single-cell barcode libraries were processed to match cell IDs from the scRNA-seq transcriptomes. Reads lacking the specified flanking regions on the U6 promoter and spacer regions ("AACAC" and "TTAGAG" respectively) or those with an Illumina Phred score of <28, were discarded (**Supplementary Fig. 5**). For each UMI with at least three barcode reads, a consensus barcode sequence was determined by selecting the sequence present in over 50% of the retained UMIs. Diversity scores were calculated as in bulk analysis by performing a pairwise alignment of each barcode with the gRNA spacer, using the pairwiseAlignment function from the Biostrings package.

Random forest model

The random forest model was employed to evaluate whether barcode editing over time could predict barcode duration in single cells. Single-cell datasets were used from days 0, 4, 6 and 10, whereby the mean barcode length, percentage of original sequences, mean diversity and coverage were incorporated as features. For each dataset, 20% of the data was set aside for testing, while the remaining 80% was used for training. The randomForest function from the randomForest package (version *4.7-1.1*) was used to fit a regression model[48], with "day" as the response variable and the selected features as predictors. Model performance was assessed by predicting day values for both the training and test sets, and the R^2 statistic was calculated to quantify the proportion of variance explained by the model.

References

1. Maška, M. *et al.* The cell tracking challenge: 10 years of objective benchmarking. *Nat. Methods* **20**, 1010–1020 (2023).

2. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

3. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).

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

5. Pickar-Oliver, A. & Gersbach, C. A. The next generation of CRISPR–Cas technologies and applications. *Nat. Rev. Mol. cell Biol.* **20**, 490–507 (2019).

6. Li, Y., Li, S., Wang, J. & Liu, G. CRISPR/Cas systems towards next-generation biosensing. *Trends Biotechnol.* **37**, 730–743 (2019).

7. Farzadfard, F. & Lu, T. K. Emerging applications for DNA writers and molecular recorders. *Science (80-.).* **361**, 870–875 (2018).

8. Bhattarai-Kline, S. *et al.* Recording gene expression order in DNA by CRISPR addition of retron barcodes. *Nature* **608**, 217–225 (2022).

9. Sheth, R. U., Yim, S. S., Wu, F. L. & Wang, H. H. Multiplex recording of cellular events over time on CRISPR biological tape. *Science (80-.).* **358**, 1457–1461 (2017).

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

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

12. Kalhor, R. *et al.* Developmental barcoding of whole mouse via homing CRISPR. *Science (80-.).* **361**, (2018).

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

14. Zhang, W. *et al.* The bone microenvironment invigorates metastatic seeds for further dissemination. *Cell*

184, 2471–2486 (2021).

15. Choi, J. *et al.* A time-resolved, multi-symbol molecular recorder via sequential genome editing. *Nature* **608**, 98–107 (2022).

16. Tang, W. & Liu, D. R. Rewritable multi-event analog recording in bacterial and mammalian cells. *Science (80-.).* **360**, eaap8992 (2018).

17. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science (80-.).* **353**, aag0511 (2016).

18. Park, J. *et al.* Recording of elapsed time and temporal information about biological events using Cas9. *Cell* **184**, 1047–1063 (2021).

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

20. McKenna, A. & Gagnon, J. A. Recording development with single cell dynamic lineage tracing. *Development* **146**, dev169730 (2019).

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

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

23. Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).

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

25. Li, L. *et al.* A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells. *Cell* **186**, 5183–5199 (2023).

26. Bowling, S. *et al.* An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells. *Cell* **181**, (2020).

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

28. Yang, D. *et al.* Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. *Cell* **185**, 1905–1923 (2022).

29. Quinn, J. J. *et al.* Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer

xenografts. *Science (80-.).* **371**, eabc1944 (2021).

30. Simeonov, K. P. *et al.* Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. *Cancer Cell* **39**, 1150–1162 (2021).

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

32. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).

33. Zuo, E. *et al.* Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science (80-.).* **364**, 289–292 (2019).

34. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).

35. Li, M. & Izpisua Belmonte, J. C. Deconstructing the pluripotency gene regulatory network. *Nat. Cell Biol.* **20**, 382–392 (2018).

36. Bošković, A. *et al.* Higher chromatin mobility supports totipotency and precedes pluripotency in vivo. *Genes Dev.* **28**, 1042–1047 (2014).

37. Eckersley-Maslin, M. A. *et al.* MERVL/Zscan4 network activation results in transient genome-wide DNA demethylation of mESCs. *Cell Rep.* **17**, 179–192 (2016).

38. Rodriguez-Terrones, D. *et al.* A molecular roadmap for the emergence of early-embryonic-like cells in culture. *Nat. Genet.* **50**, 106–119 (2018).

39. Zalzman, M. *et al.* Zscan4 regulates telomere elongation and genomic stability in ES cells. *Nature* **464**, 858–863 (2010).

40. Fu, X., Djekidel, M. N. & Zhang, Y. A transcriptional roadmap for 2C-like-to-pluripotent state transition. Sci. Adv. 6, eaay5181. (2020).

41. Beccari, L. *et al.* Multi-axial self-organization properties of mouse embryonic stem cells into gastruloids. *Nature* **562**, 272–276 (2018).

42. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).

43. Rosen, L. U. *et al.* Inter-gastruloid heterogeneity revealed by single cell transcriptomics time course: implications for organoid based perturbation studies. *bioRxiv* 2009–2022 (2022).

44. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science (80-.).* **339**, 819–823 (2013).

45. Clark, S. J. *et al.* ScNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells e. *Nat. Commun.* (2018) doi:10.1038/s41467-018-03149-4.

46. Pages, H. *et al.* Package 'Biostrings'. *Bioconductor* **18129**, (2013).

47. Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).

48. Liaw, A. & Wiener, M. Title Breiman and Cutler's Random Forests for Classification and Regression Description Classification and regres-sion based on a forest of trees using random inputs. (2015).