nature protocols

https://doi.org/10.1038/s41596-025-01141-z

Protocol

Check for updates

DARLIN mouse for in vivo lineage tracing at high efficiency and clonal diversity

Li Li[1,2,7], Sarah Bowling[3,7], Hongying Lin[2], Daolong Chen[1,2], Shou-Wen Wang[1,2,4] ✉ & Fernando D. Camargo[5,6] ✉

Abstract

Lineage tracing is a powerful tool to study cell history and cell dynamics during tissue development and homeostasis. An increasingly popular approach for lineage tracing is to generate high-frequent mutations at given genomic loci, which can serve as genetic barcodes to label different cell lineages. However, current lineage tracing mouse models suffer from low barcode diversity and limited single-cell lineage coverage. We recently developed the DARLIN mouse model by incorporating three barcoding arrays within defined genomic loci and combining Cas9 and terminal deoxynucleotidyl transferase (TdT) to improve editing diversity in each barcode array. We estimated that DARLIN generates 10^{18} distinct lineage barcodes in theory, and enables the recovery of lineage barcodes in over 70% of cells in single-cell assays. In addition, DARLIN can be induced with doxycycline to generate stable lineage barcodes across different tissues at a defined stage. Here we provide a step-by-step protocol on applying the DARLIN system for in vivo lineage tracing, including barcode induction, estimation of induction efficiency, barcode analysis with bulk and single-cell sequencing, and computational analysis. The execution time of this protocol is ~1 week for experimental data collection and ~1 d for running the computational analysis pipeline. To execute this protocol, one should be familiar with sequencing library generation and Linux operation. DARLIN opens the door to study the lineage relationships and the underlying molecular regulations across various tissues at physiological context.

Key points

- The DARLIN mouse enables the study of the cell lineages of millions of cells and at a high efficiency in vivo.

- Compared with other lineage-tracing mouse models, which can suffer from low barcode diversity and limited single-cell lineage coverage, the DARLIN mouse incorporates three barcoding arrays within defined genomic loci and combines Cas9 and terminal deoxynucleotidyl transferase to improve editing diversity in each barcode array.

Key references

Li, L. et al. *Cell* **186**, 5183–5199.e22 (2023): https://doi.org/10.1016/j.cell.2023.09.019

Bowling, S. et al. *Cell* **181**, 1410–1422.e27 (2020): https://doi.org/10.1016/j.cell.2020.04.048

Patel, S. H. et al. *Nature* **606**, 747–753 (2022): https://doi.org/10.1038/s41586-022-04804-z

[1]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China. [2]School of Life Sciences, Westlake University, Hangzhou, China. [3]Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA, USA. [4]School of Science, Westlake University, Hangzhou, China. [5]Stem Cell Program, Boston Children's Hospital, Boston, MA, USA. [6]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. [7]These authors contributed equally: Li Li, Sarah Bowling. ✉e-mail: wangshouwen@westlake.edu.cn; Fernando.Camargo@childrens.harvard.edu

Introduction

The ability to track cellular lineages alongside gene expression readouts is critical for understanding cellular behavior during development, adulthood and disease. As such, intense efforts over the past decades have been invested into developing tools that enable the tracing of cells in tissues in vivo. While early work used injectable dyes to trace cells in ex vivo systems, advances in genetic models has enabled tracing through recombinase approaches, whereby cells can be traced with labels such as fluorescent proteins[1–4]. Thanks to these approaches, major advances have been made in our understanding of organismal development, stem cell dynamics during tissue homeostasis and the steps leading to disease onset and progression[5,6]. However, the major drawback of these methods is that they suffer from low lineage-labeling diversity, which enables only a handful of cell lineages to be tracked at once. Therefore, our understanding of complex tissue dynamics and interactions within diverse cellular environments remains limited, highlighting the need for newer, more sophisticated methods that can simultaneously track a greater number of cell lineages with higher resolution and accuracy.

Development of the protocol

To address these gaps, DNA barcoding tools have recently been developed that enable high-resolution tracing of cell lineages. DNA barcoding takes advantage of the enormous diversity of information that can be both stored in DNA, allowing many cells to be labeled simultaneously, and also readout via next generation sequencing, enabling high-throughput and efficient analysis of clonal information. In the past decade, numerous *in vivo* DNA barcoding systems have been generated in zebrafish[7–10], mice[11–17] and *Drosophila*[18]. Inspired by the GESTALT system for DNA barcoding in zebrafish[7,8], we have previously generated an in vivo mouse model named Cas9–CARLIN that use CRISPR–Cas9 gene editing to produce highly diverse barcodes that act as unique and heritable markers of cells and their progeny[19]. Cas9–CARLIN contains a single expressed target array consisting of ten tandem CRISPR target sites in the *Col1a1* locus (Col1a1 array, or CA), and doxycycline (Dox)-inducible Cas9 expression enables temporal control of barcode generation at any time during development or adulthood (Fig. 1). Furthermore, the DNA barcodes are expressed, enabling their interrogation at the single-cell level alongside whole transcriptional readouts using commercially available single-cell RNA-sequencing (scRNA-seq) platforms. However, Cas9–CARLIN generates only ~44,000 distinct lineage barcodes and suffers from limited efficiency of both barcode editing and capture.

To enable the study of cell lineages of millions of cells and at a high efficiency, we developed DARLIN (Fig. 2), a substantially improved lineage tracing mouse model over Cas9–CARLIN[20]. One of the main reasons for low barcode diversity in Cas9–CARLIN is the extensive deletions introduced by Cas9 editing. To address this problem, we fused Cas9 with terminal deoxynucleotidyl transferase (TdT) so that more insertions can be added upon Cas9-induced double-strand break at the target site. Apart from increased insertions, we also observed fewer deletions in Cas9–TdT compared with Cas9 alone. Combined, Cas9–TdT improves the barcode diversity to approximately 10^6 per target array. Furthermore, we included another two target arrays (TA, target array in Tigre locus, and RA, target array in Rosa26 locus) in the DARLIN system, which reuse the ten target sites from CA but with different orders. This leads to an

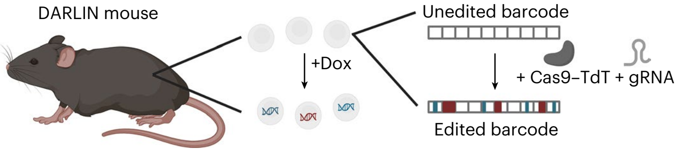


Fig. 1 | DARLIN mouse barcoding system. Dox administration results in expression of Cas9–TdT, which induces double-strand breaks in three transcribed barcode target arrays. The breaks are repaired in an error-prone manner, resulting in the generation of a diverse set of indels in the barcode sequence. These act as a unique and heritable marker of each clone. The TdT polymerase favors insertions during DNA repair, resulting in highly diverse barcode combinations. Created with BioRender.

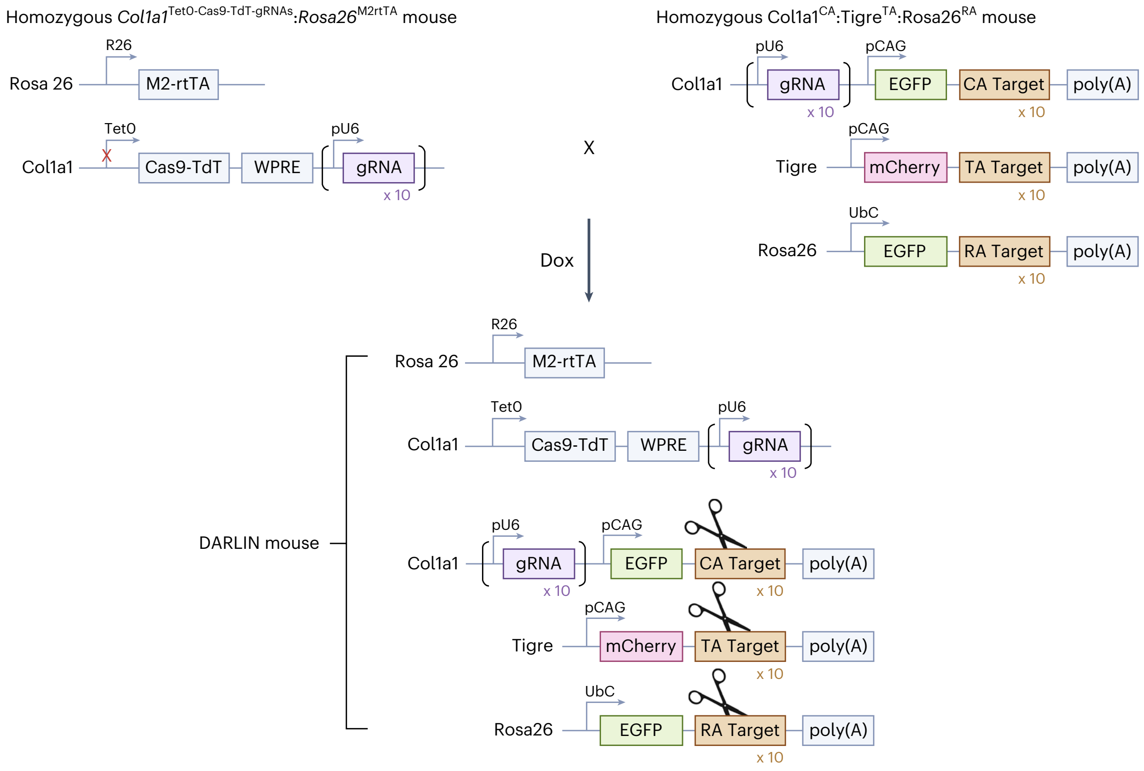


Fig. 2 | Schematic of mouse line genetics and the generation of DARLIN. The homozygous *Col1a1*$^{\text{tetO-Cas9-TdT-gRNAs}}$*Rosa26*$^{\text{M2-rtTA}}$ mouse and *Col1a1*$^{\text{CA}}$*Tigre*$^{\text{TA}}$*Rosa26*$^{\text{RA}}$ mouse are maintained separately. These two mice contain the same gRNA cassette, and each cassette contains 10 gRNAs with different sequences that match the 10 target sites in each of the three target arrays (CA, TA and RA). These 10 target sites are shuffled to have different orders in each target array. These two mouse lines are crossed to generate the DARLIN mouse, which can be induced for whole-body cell barcoding with Dox administration. M2-rtTA, the mutant tetracycline reverse transactivator; WPRE, woodchuck hepatitis virus posttranscriptional regulatory element. Created with BioRender.

estimated 10^{18} lineage barcodes in combination, which is much larger than the total cell number in an adult mouse (10^{10}). Due to the improved design, we consistently observed ~90% editing efficiency in embryos in our recent work[20], although the editing in the adult stage was more variable. In the three target arrays, CA should have the same expression as in the Cas9–CARLIN system due to the identical design. We observed that TA expression was ~3 times as high as the CA expression, while RA expression is comparable with CA expression. This high expression greatly improves the barcode capture. Overall, we observed that DARLIN enables the recovery of 'edited' lineage barcodes in over 70% cells in single-cell assays when combining measurements in all three target arrays. Individually, we observed a recovery efficiency of ~25%, ~55% and ~25% in single cells from CA, TA and RA, respectively. In DARLIN, all barcoding elements are contained within defined genomic loci, making the mouse colony maintenance straightforward (Fig. 2).

DARLIN is a retrospective lineage tracing tool that nonspecifically labels cells across the whole body. These cells are profiled at typically a single timepoint for their lineage identity and other molecular profiles. The resulting lineage tracing data require more sophisticated computational analysis and interpretation. We have previously developed CoSpar to facilitate exploratory analysis of clonal barcoding data[21]. Applying CoSpar to DARLIN data, we have successfully inferred the differentiation hierarchy of hematopoiesis and estimated the migration dynamics of hematopoietic stem cells (HSCs)[20]. With the fate bias prediction from CoSpar, we also identified the subpopulation of HSCs that are biased toward megakaryocytes[20]. There is still a strong need for computational tools that can extract more information from such lineage tracing data, such as inferring differentiation dynamics or the phylogenetic relationships[22–24].

Applications

So far, barcoding-based in vivo lineage tracing systems such as DARLIN have been applied to study diverse areas of stem cell biology, including hematopoietic cell migration[19,20], lineage branching points during development[19,25], hematopoiesis developmental hierarchy[13,14,20], epigenetic memory of HSCs[20], HSC and T cell fate biases[20,26–29] and HSC responses to chemotherapy[19]. Since the target arrays in DARLIN are contained within well-described safe-harbor loci, the mutations in these loci probably have negligible effects on cell dynamics in mice. Since DARLIN could label a wide range of tissues, it can be applied to study development, adult tissue homeostasis or disease state in other tissues such as lung, liver, kidney and so on.

DARLIN can be combined with other tools to enable more powerful applications for lineage tracing. First, dissecting cell fate choice is a key application of lineage tracing tools, and many mouse models have been developed to specifically label a desired population with fluorescent markers to study their fate choices over time[1–4]. These tools cannot resolve the heterogeneous fate choices within the labeled population. One interesting direction would be to combine the massive barcoding diversity of DARLIN with the specific labeling of other mouse lines to enable high-resolution study of lineage and fate choices within a specific population of cells. Second, DARLIN can be utilized for single-cell multiomic lineage tracing, as epigenomic information plays important roles in regulating lineage dynamics and cell identity. We have demonstrated this with Camellia-seq, a method that we developed along with DARLIN to simultaneously profiles DNA methylation, gene expression, chromatin accessibility and lineage barcodes in single cells[20]. Applying Camellia-seq to DARLIN leads to the observation of much stronger epigenetic memory in DNA methylation than in gene expression or chromatin accessibility. Camellia-seq is a low-throughput protocol. To increase the throughput, it is possible to adapt the commercially available multiome kit from 10X genomics to jointly profile gene expression, chromatin accessibility and lineage barcodes from single cells[30]. Finally, current single-cell lineage tracing lacks the spatial information of these cells, which could be critical for understanding the dynamics and fate choices of individual cells[31]. Spatial transcriptomics methods have been substantially developed to enable higher resolution and throughput of transcriptomics in situ[32–34]. It will be an exciting opportunity and challenge to integrate DARLIN with spatial transcriptomics and generate spatially resolved clonal information directly on intact tissues.

Comparison with other methods

Existing in vivo lineage-barcoding mouse models use random transposon insertion[14], Cre-lox recombination[13,15,16] or CRISPR scarring[11,12] to induce DNA barcodes. The ability of joint lineage barcode and transcriptome profiling is important. However, transposon-based barcoding models such as Sleeping Beauty lack expressed barcodes[14]. Although the barcodes from Cre-lox recombination such as PolyloxExpress can be expressed as mRNA, these barcodes have thousands of base pairs and thus require more specialized and expensive sequencing approaches[13,15]. In addition, the detection efficiency of these barcodes is also limited[15]. Finally, although the CRISPR-based MARC1 mouse has enormous barcoding diversity due to its 60 homing guides across the genome[12], these 60 homing guides are integrated randomly into both the maternal and paternal genome, which makes it difficult to maintain the line and may perturb the native biology in mice. More importantly, these barcodes are not expressed as mRNA, thus difficult to measure in single cells with standard sequencing approaches.

In contrast, DARLIN barcoding elements are placed in three well-defined safe loci to enable simple maintenance and mitigate possible perturbations on native cell dynamics. Furthermore, DARLIN barcodes are relatively short (~270 bp), and are highly expressed as mRNA so that they can be readout in single cells with standard Illumina sequencing, jointly with their transcriptomic and/or epigenomic measurements. Importantly, DARLIN features a massive lineage barcode diversity (~10^6 per target array, or 10^{18} when combining three target arrays), high editing efficiency (~100% when induction happens in the embryos and less for adult induction), and highly expressed barcodes (especially for TA, ~3 times that of CA or RA). In comparison, Cas9–CARLIN generates only ~44,000 barcodes, labels cells with limited efficiency (16–74%), and has only a single expressed barcode (that is, CA). We estimated in DARLIN that we can recover at least one edited lineage barcodes from either CA, TA or RA in ~70% of profiled single cells,

Table 1 | Comparison of CARLIN and DARLIN barcoding systems

		CARLIN	DARLIN
Editing efficiency	Embryo induction	~20%	~90%
	Adult induction	16–88%	18–100%
Number of barcode arrays		1	3
Editing protein		Cas9	Cas9–TdT
Additional features		NA	WPRE (increase Cas9–TdT expression)
Capture efficiency for both edited and unedited barcodes	Col1a1 locus	33–67%	25–35%
	Tigre locus	NA	52–73%
	Rosa26 locus	NA	25–40%
	All loci (cell frac. with ≥1 barcode detected)	NA	~80%
Max. barcode diversity	Using one BC array	44,000	~10^{6}
	Using three BC arrays	NA	~10^{18}

NA, not applicable. Numbers are reported as described in Bowling et al., 2020[19] and Li et al., 2023[20].

compared with ~15% in Cas9–CARLIN. DARLIN also compares favorably with PolyloxExpress, which reports up to theoretically 10^{6} barcoding diversity and achieves only a few percentages of single-cell lineage coverage[13,15]. Taken together, we believe that DARLIN will be the preferred barcoding mouse models for most lineage tracing studies in mice (see Table 1 for more details).

For certain tissues, it is possible that lineage barcoding in DARLIN could be inefficient (Table 2). For example, in adult brain, we previously observed that Cas9–CARLIN failed to induce barcode editing upon Dox induction in these loci. This is also probably the case for DARLIN since they share similar design in Cas9 induction. We expect that DARLIN could still be used to study brain development when barcoding happens at the embryonic stage, and other target arrays such as TA and RA could be profiled for lineage information. However, to study lineage dynamics in adult brain or other inaccessible tissues in DARLIN, alternative mouse models that could work in these tissues would be preferred[17].

Finally, careful experimental design is needed to study adaptive immune cells with DARLIN. TdT is expressed in T and B cells to diversify the immune repertoire. Transient expression of Cas9–TdT in immune cells may perturb their dynamics over a short period of time. Therefore, lineage profiling should occur long after Dox induction to allow the immune system to recover from such perturbation. Alternatively, other expressible barcoding systems, though with less barcode diversity, could be used in such studies. For example, Dox-inducible Cas9 mice can be crossed with the homozygous *Col1a1*^{CA}*Tigre*^{TA}*Rosa26*^{RA} mouse to study the immune system to minimize TdT-induced perturbation.

Overview of the procedure

Below, we discuss the considerations needed for a successful lineage barcoding experiment, DARLIN mouse line maintenance and procedures for lineage tracing with DARLIN. Depending on the biological questions to be addressed, one could choose to barcode the cells in DARLIN at either the embryonic or adult stages (Step 1), and profile the barcodes using either bulk

Table 2 | Summary of editing efficiency across different tissues in CARLIN and DARLIN mouse upon Dox induction at adult stage

	Intestine	Kidney	Liver	Lung	Gonad	Spleen	Skin	Brain	Muscle	Heart
CARLIN	~45%	~30%	~45%	~40%	~90%	~40%	~50%	~1%	0%	0%
DARLIN	100%	100%	~95%	~99%	~99%	~99%	NA	NA	NA	NA

NA, not applicable. The CARLIN data are based on Fig. 3b from Bowling & Sritharan et al.[19], while the DARLIN data are from Fig. 2j from Li et al.[20]. Although certain tissues may not be edited when induced in the adult stage, we expect to see high editing with embryonic induction. Note that the editing data in CARLIN are based on a single mouse, and the data in DARLIN are from three mice with the highest editing. The exact editing efficiency may vary between mice.

(Steps 6–23) or single-cell (Steps 24–37) sequencing at a later stage. After Dox induction, we recommend roughly estimating the Dox editing efficiency (Steps 2–5) to ensure a successful barcoding experiment. This can be done before tissue dissection for sequencing. The sequencing data are then preprocessed separately depending on whether they are bulk or single-cell data (Steps 38–43). Downstream analysis after data preprocessing needs more customization and will be covered in our GitHub repository (https://github.com/ShouWenWang-Lab/DARLIN_tutorial).

The single-cell sequencing approach can be used to jointly profile both the clonal and transcriptomic information, enabling a more systematic study of cell lineage dynamics. Although the bulk assay does not associate each clonal (DARLIN) barcode at the single-cell resolution, it enables to study the clonal composition in a much larger population in a cost-effective way. Furthermore, when a given cell type of interest has highly specific purification markers for fluorescence-activated cell sorting (FACS), as in the case of hematopoiesis, one can still have both the clonal identity and phenotypic information with the bulk assay, as we demonstrated when studying HSC migration in our recent work[20].

Experimental design

Considerations of lineage barcoding experiments

To carry out a successful lineage tracing experiment, it is important to obtain enough large clones (Fig. 3). While a higher editing efficiency leads to more clones labeled (Fig. 3b), it is often not necessary to label every cell at the time of induction. Assume that there are N progenitors at the time of barcoding (which can be estimated from a pilot experiment), a fraction φ of them have a fate bias toward a certain cell type, and the barcode editing efficiency is η. To convincingly identify such a fate bias, one should observe at least ~10 fate-biased clones, i.e., $N\eta\varphi \geq 10$. Therefore, the required editing efficiency in this case is:

$$\eta \geq \frac{10}{N\phi}.$$

Therefore, a smaller editing efficiency would be sufficient for a larger initial progenitor population that have stronger fate bias toward a certain cell type. However, when proving that certain stem cells do not generate a given fate outcome or do not differentiate at all, for example, to prove that hematopoietic stem cells do not actively produce mature blood cell types in adult unperturbed mice, one would desire nearly 100% editing efficiency to avoid missing such a differentiation behavior. The editing efficiency can be modulated with the Dox concentration. Due to variabilities in Dox induction, we would suggest carrying out barcode labeling in several DARLIN mice simultaneously and estimate editing efficiency in each mouse (Steps 3–6) to decide the ones that have sufficient editing for a successful lineage tracing experiment. The actual lineage tracing datasets should be generated from such qualified mice.

In addition, sampling more cells is associated with larger observed clones (Fig. 3c). To be useful in downstream clonal analysis, these clones need to have at least two cells detected at the time of observation. Assume that the average size of these clones is λ after clonal expansion during the tracing period (Fig. 3a) and that a fraction p of the entire desired population is

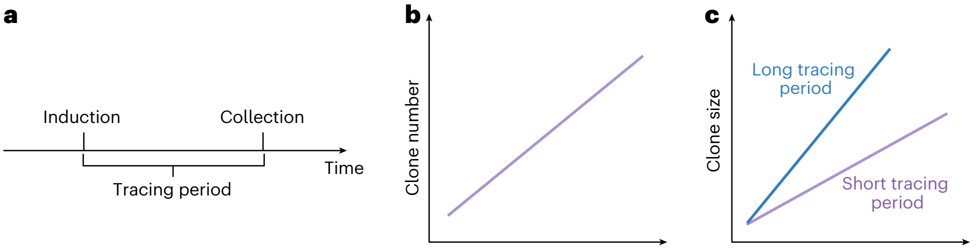


Fig. 3 | General considerations for lineage tracing experiments. a, Experiment diagram of lineage tracing in DARLIN mice. **b**, The relationship between barcode editing efficiency and clone number. **c**, The relationship between cell sample fraction and clone size with different tracing periods.

captured and sequenced at the time of collection. Then, to ensure that each sampled clone has on average ≥2 cells, the minimum sampling fraction should be:

$$p \geq \frac{2}{\lambda}.$$

Here, λ can be estimated in a pilot experiment. Note that we are not trying to make claims based on a single clone with just two cells. Instead, we are aiming to infer collective patterns across many clones with ≥2 cells, which would be statistically significant. Sometimes, even this minimum fraction could result in millions of cells for sequencing, which is costly. Fortunately, the clonal expansion is heterogeneous, and some clones can be much larger than others. Using a much smaller sampling fraction may still capture enough usable clones with ≥2 cells. In such cases, a high barcode labeling efficiency also leads to a higher chance of capturing enough usable clones when just sequencing a very small fraction of cells in the tissue. Ideally, a realistic simulation involving stochastic division and differentiation can be performed with input parameters estimated from a pilot experiment to determine the optimal sample size that balance the cost and desired number of informative clones. To do this, one may start with the function cospar.simulate.bifurcation_model provided in our CoSpar package (https://github.com/ShouWenWang-Lab/cospar/blob/master/cospar/simulate.py#L258), and modify it for your specific experiments.

Lineage barcode induction and efficiency estimation in DARLIN (Steps 1–5)

The DARLIN mouse model is generated by crossing homozygous *Col1a1*$^{\text{tetO-Cas9-TdT-gRNA/tetO-Cas9-TdT-gRNA}}$:*Rosa26*$^{\text{M2-rtTA/M2-rtTA}}$ mouse and homozygous *Col1a1*$^{\text{CA/CA}}$:*Tigre*$^{\text{TA/TA}}$:*Rosa26*$^{\text{RA/RA}}$ mouse. To avoid background editing of DARLIN arrays, we maintain Cas9-TdT-gRNAs-M2 and CA/TA/RA mouse lines separately in homozygosity until DARLIN is needed (Fig. 2).

The barcoding efficiency of the DARLIN system depends on the concentration and duration of Dox treatment. To induce lineage barcodes in DARLIN embryos, timed pregnancies are setup and 50 µg/g Dox is injected at the desired embryonic stage into the pregnant dam through a retro-orbital route. To generate barcodes in neonatal or adult DARLIN mice, Dox is administered via drinking water for 1 week (2 mg/ml, supplemented with 10 mg/ml sucrose) and additionally with three intraperitoneal injections (50 µg/g) every other day (days 1, 3, 5 of water administration) during the same week. Before performing bulk or single-cell analysis with DARLIN, we highly recommend estimating the labeling efficiency of the DARLIN barcodes. Although the genomic DNA from the tissue of interest would be ideal for estimating the editing efficiency, mouse tail can be substituted for rough estimation, which can be extracted conveniently from a live mouse. We extract genomic DNA from a small amount of mouse tissue or the tip of mouse tail, amplify CA, TA and RA, respectively, from the genomic DNA, and run an agarose gel to evaluate the editing efficiency of each DARLIN array (Fig. 4). As it is shown in Fig. 4, the editing efficiency of triple DARLIN array in mice 1–3 is nearly 100%, and it is a bit lower in mouse 4.

Bulk DARLIN library preparation (Steps 6–23)

Bulk DARLIN analysis provides a method for obtaining information about the lineage relationships between different tissues or cell types[20]. It is particularly helpful when phenotypic

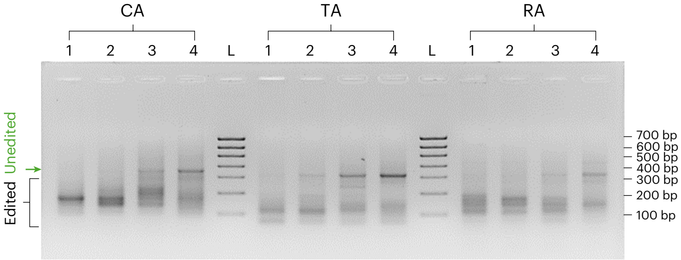


Fig. 4 | Representative agarose gel image of edited DARLIN array. Dox was injected in four DARLIN mice at E10. At 3 weeks old, the editing efficiency in each mouse was estimated from the mouse tail tip, by running an agarose gel of the amplification product from CA, TA and RA, respectively. The labels 1–4 represent the mouse ID, and L represents DNA Ladder, whose reference DNA lengths are marked on the right.

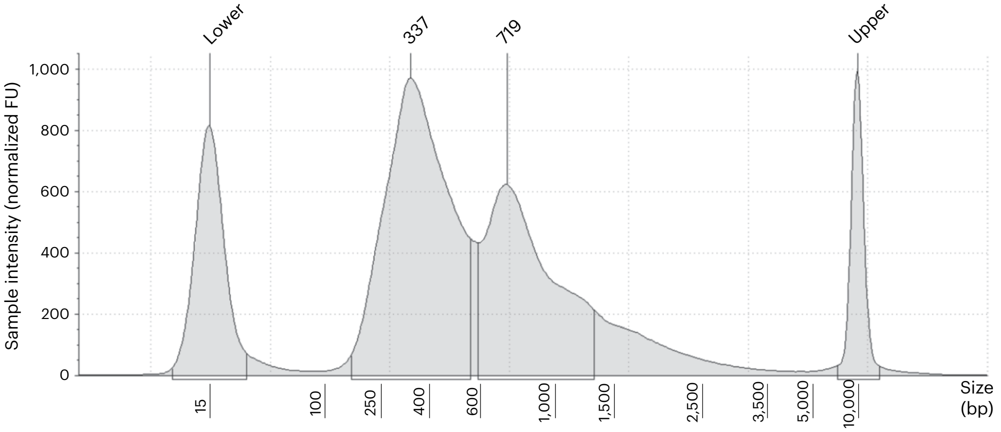


Fig. 5 | Representative TapeStation results of bulk DARLIN libraries. Bulk DARLIN library of TA was generated with the tail tip of mouse 1 in Fig. 4. The library was run with TapeStation High Sensitivity D5000 reagent. Lower (15 bp) and upper (10,000 bp) ladders are shown in the plot. The major peak at ~350 bp corresponds to the DARLIN sequences. Fragments longer than 600 bp may arise from unspecific PCR amplification. In our experience, these undesired fragments have a negligible impact on the final ratio of valid DARLIN reads, probably because shorter fragments bind more efficiently on the flow cell. SPRI beads can be used to exclude undesired long fragments.

information can be obtained jointly using FACS purification, as discussed above. Although it is feasible to perform bulk DARLIN analysis from DNA, we recommend to amplify from mRNA as there is only one copy of the barcode in the DNA but many more in mRNA due to high expression of the DARLIN barcodes. Different commercial kits for RNA extraction and purification have variable requirements for the minimal cell numbers. Here, we use TRIzol to standardize the protocol and accommodate a variety of cell numbers (10^2–10^6) in these experiments. Next, we perform reverse transcription with a cocktail of annealing primers for CA, TA and RA arrays to amplify all three kinds of DARLIN arrays from the same sample. Specifically, we utilize a nested PCR approach to ensure the specificity of this amplification. In addition, while we perform the first round of PCR amplification with mixed CA, TA and RA primers, in later rounds we separately amplify each DARLIN array using the corresponding primer set. This can minimize the amplification bias due to the initial RNA expression differences between these arrays. We use 1.5× AMPure XP beads (Beckman Coulter, A63881) to purify the PCR product from each step except the final indexing PCR product (0.8×), since the edited DARLIN array ranges between 50 bp and 300 bp. Finally, the DARLIN library is analyzed with a TapeStation (Fig. 5) and sequenced on an Illumina MiSeq using a paired-end 500-cycle kit.

DARLIN library preparation from scRNA-seq libraries (Steps 24–37)

It is also possible to amplify DARLIN barcodes from scRNA-seq libraries, enabling the clonal relationship of cells to be analyzed alongside transcriptional readouts. Here, we detail the steps for amplifying DARLIN barcodes from scRNA-seq libraries generated by the commercially available 10X Genomics Chromium Single Cell 3′ reagent kits. The manufacturer's protocol for encapsulating the cells, performing reverse transcription and amplifying cDNA is followed up until step 2.4, when DARLIN barcodes are separately amplified from the library. Similar to bulk amplification, we use a nested PCR approach to minimize amplification bias of shorter edited fragments and to boost barcode capture. DARLIN single-cell libraries are sequenced on an Illumina MiSeq. When using plate-based STRT-seq to profile single cell transcriptome, we observed similar efficiency of barcode recovery. We also expected similar or better results from SMART-seq.

Analysis of lineage tracing data from DARLIN (Steps 38–43)

We have also developed computational methods to facilitate the preprocessing of raw sequencing data and the identification of reliable clones from DARLIN. A MATLAB-based CARLIN pipeline was developed initially to analyze data from Cas9–CARLIN. Building on

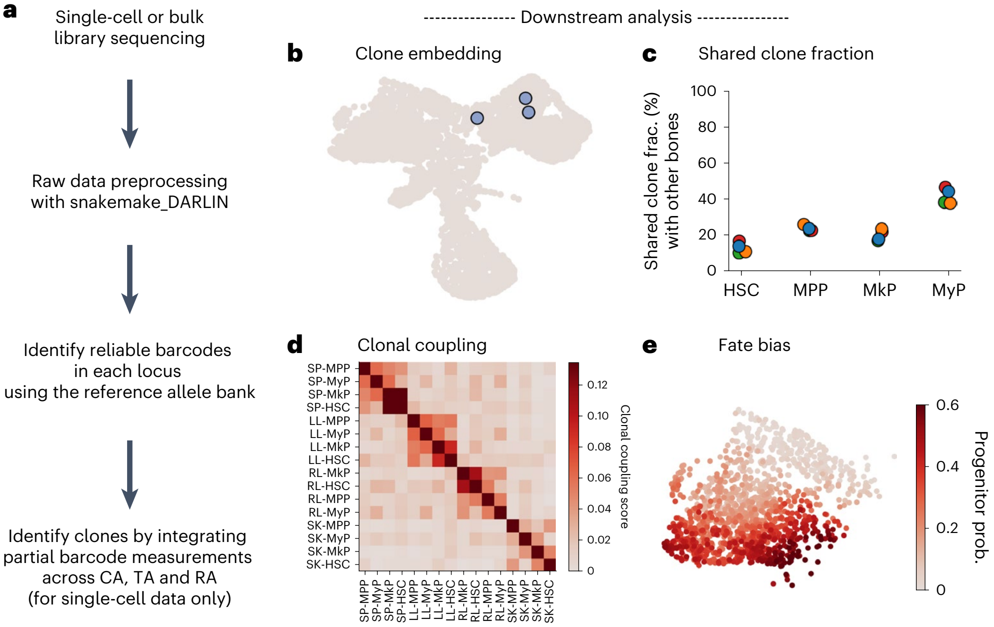
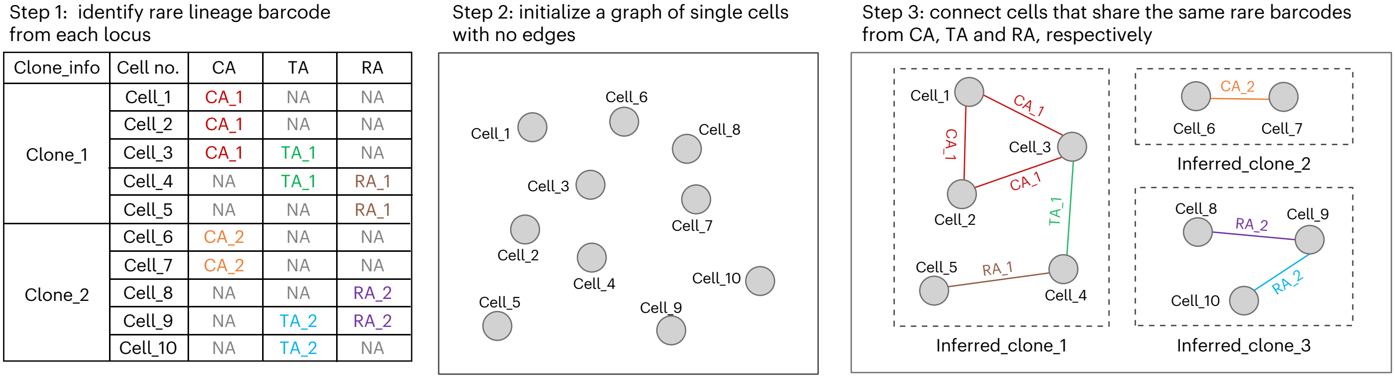


Fig. 6 | DARLIN data analysis. a, Workflow of DARLIN data analysis. After sequencing, the fastq data are processed with snakemake_DARLIN pipeline to extract lineage barcodes. Common lineage barcodes may be generated simultaneously in multiple unrelated cells, leading to barcode homoplasy. These barcodes are excluded after querying the reference allele bank. Finally, for single-cell lineage tracing data that jointly profiles CA, TA and RA in the same cell, the barcode information are integrated across these loci to jointly identify cells that come from the same clone. While the data preprocessing with snakemake_DARLIN is covered in great details in this procedure (Steps 38–42), the remaining analyses are only briefly mentioned in Step 43 and can be found in our online tutorial (https://github.com/ShouWenWang-Lab/DARLIN_tutorial). **b**–**e**, examples of downstream analysis: clone visualization on single-cell transcriptomic embedding (**b**), quantification of shared clones between different cell populations (**c**), lineage coupling calculation between different cell populations (**d**) and fate bias prediction with CoSpar (**e**). Images **c** and **e** are reproduced permission from ref. 20, Elsevier. MPP, multipotent progenitor; MkP, megakaryocyte progenitor; MyP, myeloid progenitor; frac., fraction; prob., probability.

this, we developed a pipeline named snakemake_DARLIN[20] that conveniently takes care of data from different sequencing types and target loci, and enables reproducible and parallel preprocessing of multiple sequencing samples in a Linux-based high-performance computer cluster. A schematic of our analysis workflow is illustrated in Fig. 6a. The data preprocessing is standard and is covered in the following procedures and discussed in Steps 38–42. However, further data analyses such as barcode filtering, clone identification and clonal relationship estimation are more complex and only briefly mentioned in the following procedures (Step 43). Below, we discuss these additional data analyses that are covered in our online tutorial (https://github.com/ShouWenWang-Lab/DARLIN_tutorial).

Like most in vivo barcoding systems, DARLIN generates both frequent and infrequent lineage barcodes. Most of the lineage barcodes occur at very low frequencies, which are reliable to uniquely label individual clones. However, some mutation patterns or barcodes could be generated concurrently in several cells from unrelated lineages, leading to barcode homoplasy. To address this challenge, we have previously collected a relatively large reference dataset with estimated intrinsic generation probability for ~10^5 lineage barcodes in each of the three target loci (Fig. 2). Since we expect ~10^6 lineage barcodes per locus, this allele bank is far from exhaustion. However, we expect that most of the high-frequency lineage barcodes are included in this reference dataset. The reference dataset is provided in our MosaicLineage package[20] (https://github.com/ShouWenWang-Lab/MosaicLineage/tree/master/reference). In a lineage-tracing study, observed lineage barcodes can be queried in our reference dataset and common barcodes above an appropriate cutoff of generation probability can be excluded. This probability threshold depends on the expected number of clones to be present in the dataset. Barcodes generated at a relatively high probability can still provide unique labels if there are only a few clones expected. Following our previous mathematical derivation[20], we recommend excluding alleles with a barcode generation probability ρ higher than ρ^*, where ρ^* is the probability cutoff determined from $\langle \rho | \rho \leq \rho^* \rangle = 2\alpha/(M-1)$. Here, $\langle \rho | \rho \leq \rho^* \rangle$ is the average generation probability of observed barcodes below the probability cutoff ρ^*, α is the desired false discovery rate and M is the total number of expected clones in the data, which can be approximated by the observed barcode number. Roughly, we have $\rho^* \sim \alpha/M$. Although we recommend using $\alpha = 0.01$, users may set their preferred value. If one is only interested in large clones above a pre-set clone size threshold, then one may use a smaller M corresponding to only these large-clone candidates. Since our reference allele dataset is far from exhaustive, additional precaution may be taken to further mitigate barcode homoplasy, if this is desired. For example, we only used de novo barcodes (that is, barcodes not observed in our reference dataset) that also have more complex insertion patterns to infer HSC migration[20]. However, this additional filtering would also reduce the number of useful clones in a dataset.

Since DARLIN has three target arrays and among cells with detected lineage barcodes only ~10% have detected barcodes across all three loci, how to reliably identify cells that belong to the same clone from these partially observed clonal data is a great challenge. This is further complicated by the fact that some barcodes from given genomic loci have higher generation probability than others, which may lead to barcode homoplasy. We only use rare lineage barcodes detected in these genomic loci that pass the above filtering to infer clonal relationship, which partially addresses the challenge of barcode homoplasy. To illustrate the problem of incomplete barcode measurement, let us consider the following example. Cells 1–5 constitutes a clone that originates from the same founder cell. However, we only detected the CA barcode in cells 1–3, TA barcode from cells 3 and 4, and RA barcode from cells 4 and 5. Therefore, according to CA barcode, cells 1 and 2 belong to the same clone; using TA barcode, cells 3 and 4 come from the same clone; and with RA barcode, cells 4 and 5 are clonally related. Integrating information from all three arrays, we conclude that cells 1 and 2 are from one clone, while cells 3–5 are from another. In this simple example, we see that using a single target array identifies a small fraction of cells in a clone due to subsampling. Integrating information across loci help to aggregate small clones identified with each locus to generate larger clones, although some subclones may fail to be aggregated together due to missing shared cells. The aggregated bigger clones can be more informative and generate more robust conclusions. To partially address this challenge of incomplete measurement, we converted the single-cell clonal data into a network, where each cell was a vertex and two vertexes were connected if they share a rare lineage barcode from one of the three target arrays (Fig. 7). Reliable clones were identified using appropriate clustering approaches. This simplified method is provided in the function 'MosaicLineage.DARLIN.assign_clone_id_by_integrating_locus' from our MosaicLineage package. Although clone identification is not covered in the following procedures, a related tutorial is provided in https://github.com/ShouWenWang-Lab/DARLIN_tutorial/blob/master/Single-cell-tutorial-Part_1_clone_calling.ipynb. We are working on further improvement to solve this problem more systematically.



Clone_info	Cell no.	CA	TA	RA
Clone_1	Cell_1	CA_1	NA	NA
	Cell_2	CA_1	NA	NA
	Cell_3	CA_1	TA_1	NA
	Cell_4	NA	TA_1	RA_1
	Cell_5	NA	NA	RA_1
Clone_2	Cell_6	CA_2	NA	NA
	Cell_7	CA_2	NA	NA
	Cell_8	NA	NA	RA_2
	Cell_9	NA	TA_2	RA_2
	Cell_10	NA	TA_2	NA

Fig. 7 | Clone identification from partial barcode measurements in the three genomic loci in DARLIN. Schematic showing the three steps required for clone identification from partial barcode measurements.

It is possible to use DARLIN to continuously track cell lineages over many cell divisions. In our system, we have 30 target sites across all three target arrays. In principle, if any of these target sites is edited, it is unlikely to be edited again due to mismatch between this target site and the corresponding gRNA. Therefore, the shared mutations between cells could be utilized to build a lineage tree reflecting the cell division histories. However, there could be large deletions that erase prior mutations within the deleted region, adding computational challenges for tree inference. Although we have not carried out such a study with DARLIN, we have the following suggestions for interested users. First, to achieve continuous editing over a longer timescale, we suggest using a much lower Dox concentration than the one we used for one-pulse barcoding. Second, we suggest using only the rare and shared mutations in this data for lineage tree reconstruction, while also considering the possible missing mutations from large deletions. The generation probability of individual mutations can be estimated from the provided allele bank dataset. Lineage tree inference from CRISPR–Cas9 editing with missing mutation information has been considered previously[19,23], which could provide a useful starting point.

Expertise needed to implement the protocol

The protocols described here require expertise in standard molecular biology techniques, animal husbandry and computational analysis (familiarity with high-performance computing, Unix shell and Python). During library preparation steps, extreme care should be taken to avoid cross-contamination of samples as errors will be amplified in the PCR steps. Furthermore, specialized core facilities will be required for animal housing and next generation sequencing steps.

Limitations

The lineage barcode labeling is not 100%, especially in adult induction. Building on this, the severe cell subsampling further complicates data interpretation. Therefore, careful statistical tests should be done to account for incomplete sampling when necessary. In addition, DARLIN still faces problems of large deletions that could erase intermediate mutations, which complicates the inference of cell division phylogeny from observed mutations. Furthermore, as discussed above, DARLIN likely shares similarly low barcode editing efficiency as Cas9–CARLIN in the brain, muscle and heart when induced at adult stage[19,35].

Materials

Biological materials

- Cas9-TdT-gRNAs-M2 mice (Jax stock no. 038749) and CA/TA/RA triple target-array mice (Jax stock no. 038750) can be ordered from the Jackson Laboratory. These two mice have C57BL/6J background

  ▲ **CAUTION** Any experiments involving live mice must conform to relevant institutional and national regulations. We obtained permission from Westlake University and Boston Children's Hospital.

Reagents

Barcode induction in DARLIN mice

- Dox (Sigma-Aldrich, cat. no. D9891)
- Sucrose (P212121, cat. no. CI-00811-5KG)
- Nuclease-free water (Qiagen, cat. no. 129115)
- Phire Tissue Direct PCR Master Mix (Thermo Scientific, cat. no. F170S)

RNA extraction

- TRIzol Reagent (Invitrogen, cat. no. 15596018)
- Chloroform (Sigma-Aldrich, cat. no. 366927)
- Isopropanol (Sigma-Aldrich, cat. no. 33539-M)

- Ethanol (Sigma-Aldrich, cat. no. 32205)
- Glycogen (Invitrogen, cat. no. AM9510)

Primers for DARLIN array amplification

- Primer sequences are listed in Table 3

Amplification of DARLIN from bulk RNA

- SuperScript III (Invitrogen, cat. no.18080093)
- Q5 High-Fidelity DNA Polymerase (New England Biosciences, cat. no. M0491L)
- Ampure XP beads (Beckman Coulter, cat. no. A63881)
- Index primers (New England Biosciences, cat. no. E7500S)

Amplification of DARLIN arrays from scRNA-sequencing libraries

- Chromium Single Cell 3′ reagent kits v3.1 (10X genomics, single cell kit cat. no. PN-1000268, Chip G kit cat. no. PN-1000120, dual index kit cat. no. PN-1000215)
- KAPA HiFi HotStart ReadyMix polymerase (Roche Sequencing Solutions, cat. no. 07958935001)
- Buffer EB (Qiagen, cat. no. 19086)
- SPRIselect beads (Beckman Coulter cat. no. B23318).

Library QC

- Qubit dsDNA HS assay kit (Invitrogen, cat. no. Q32854)
- TapeStation High Sensitivity D5000 ScreenTape (Agilent, cat. no. 5067-5592)
- TapeStation High Sensitivity D5000 Reagents (Agilent, cat. no. 5067-5593)
- Library quantification (Kappa Biosystems, cat. no. KK4835)

Table 3 | Primers for DARLIN Array Amplification

Primers for estimating the induction efficiency of CA/TA/RA	
CA_F	GAGCTGTACAAGTAAGCGGC
CA_R	GCAACTAGAAGGCACAGTCG
TA_F	GCTCGGTACCTCGCGAAT
TA_R	GCAACTAGAAGGCACCGACA
RA_F	ATGTACAAGTAAAGCGGCCG
RA_R	GCAACTAGAAGGCACACAGC
Primers for CA/TA/RA amplification from bulk RNA	
RT_CA_12UMI	CTACACGACGCTCTTCCGATCTNNNNNNNNNNNNGCAACTAGAAGGCACAGTCG
RT_TA_14UMI	CTACACGACGCTCTTCCGATCTNNNNNNNNNNNNNNGCAACTAGAAGGCACCGACA
RT_RA_14UMI	CTACACGACGCTCTTCCGATCTNNNNNNNNNNNNNNGCAACTAGAAGGCACACAGC
NGS_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
NGS_CA_R1	GTCCTGCTGGAGTTCGTGAC
NGS_TA_R1	GACGAGTCGGATCTCCCTTT
NGS_RA_R1	CGGGGATCCTCTAGAGTCG
NGS_CA_R2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAGCTGTACAAGTAAGCGGC
NGS_TA_R2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCTCGGTACCTCGCGAAT
NGS_RA_R2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGTACAAGTAAAGCGGCCG
Primers for CA/TA/RA amplification from 10X cDNA	
P5_PR1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC
NGS_CA_R1	GTCCTGCTGGAGTTCGTGAC
NGS_TA_R1	GACGAGTCGGATCTCCCTTT
NGS_RA_R1	CGGGGATCCTCTAGAGTCG
NGS_CA_R2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAGCTGTACAAGTAAGCGGC
NGS_TA_R2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCTCGGTACCTCGCGAAT
NGS_RA_R2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGTACAAGTAAAGCGGCCG

Library sequencing

- Illumina MiSeq paired-end 500-cycle v2 kits (Illumina, cat. no. MS-102-2003)
- PhiX sequencing control v3 (Illumina, cat. no. FC-110-3001)

Software

- snakemake_DARLIN (https://github.com/ShouWenWang-Lab/snakemake_DARLIN)
- MosaicLineage (https://github.com/ShouWenWang-Lab/MosaicLineage)
- CoSpar (https://cospar.readthedocs.io/en/latest/)
- MATLAB (https://www.mathworks.com/products/matlab.html)
- FastQC (https://github.com/s-andrews/FastQC)
- MultiQC (https://github.com/MultiQC/MultiQC)

Equipment

- Gross Anatomy Probe, angled (Fine Science Tools, cat. no. 10088-15)
- Insulin syringe (VWR, cat. no. BD-329461)
- NanoDrop microvolume spectrophotometer (Thermo Scientific, cat. no. ND-ONEC-W)
- PCR thermal cycler (Bio-Rad, cat. no. T100)
- PCR tubes (8-well PCR strips; Thermo Scientific, cat. no. AB-0266)
- Vortex mixer (Scientific Industries, cat. no. SI-0236)
- Refrigerated centrifuge for 1.5 and 2 ml tubes (Thermo Fisher Scientific, cat. no. 75002421)
- DynaMag-96 Side Magnet (Thermo Scientific, cat. no. 12331D)
- Qubit fluorometer (Invitrogen, cat. no. Q32866)
- Agilent 2200 TapeStation system (Agilent, cat. no. G2965AA)
- Illumine Miseq Instrument

Reagent setup

Dox solution

Dox solution is freshly made for each experiment. For injections, dissolve Dox powder in water at the concentration of 10 mg/ml. To make Dox drinking water, dissolve 1 g Dox powder and 5 g sucrose in 500 ml sterile water.

Equipment setup

Data processing setup

For data processing, we recommend using the snakemake_DARLIN pipeline described in our DARLIN paper, which was built upon the previous CARLIN pipeline. This is a snakemake-based pipeline that facilitates reproducible data processing in a convenient way. We recommend running the pipeline on a high-performance computing Linux cluster, because this processing is resource intensive, and can take ~2 h to process a single sample with 6 million reads. snakemake_DARLIN supports job submission through slurm, so that multiple samples can be processed separately and simultaneously. Typically, the installation time for the pipeline is 30 min, and the run time for the demo data is just a few minutes.

Here, we provide guidelines on how to set up the computational analysis environment needed to preprocess the DARLIN sequencing data using snakemake_DARLIN. For more details, we recommend to follow the README instructions on github (https://github.com/ShouWenWang-Lab/snakemake_DARLIN). First, set up the conda environment as follows:

```
kernel_name='snakemake_darlin'
conda create -n $kernel_name python=3.9 --yes
conda activate $kernel_name
conda install -c conda-forge mamba --yes
mamba install -c conda-forge -c bioconda snakemake=7.24.0 --yes
pip install --user ipykernel
pip install jupyterlab umi_tools seaborn papermill biopython
python -m ipykernel install --user --name=$kernel_name
```

Then, clone the necessary packages to the directory where you want to put the packages:

```
git clone https://github.com/ShouWenWang-Lab/snakemake_DARLIN --depth=1
cd snakemake_DARLIN
python setup.py develop
cd ..
mkdir CARLIN_pipeline
cd CARLIN_pipeline
git clone https://github.com/ShouWenWang-Lab/Custom_CARLIN --depth=1
```

Next, install pear (https://www.h-its.org/downloads/pear-academic/), which is used for merging read 1 and 2 when processing the bulk DARLIN data. Also, install MATLAB so that it is available directly in the command line interface or can be loaded with the command 'module load matlab'. In addition, MATLAB should have Bioinformatics Toolbox and Image Processing Toolbox addons installed. Please also install FastQC and MultiQC so that they are also available from the command line. They are not essential for data processing and the pipeline finishes properly without them. However, an informative quality report of fastq files can be obtained after running the pipeline with these two tools installed. Finally, run the test module under snakemake_DARLIN to test whether the pipeline is installed correctly.

```
cd snakemake_DARLIN/test
bash test.sh
```

The expected output can be found at the README of the github page. This test folder includes three datasets of different types and their corresponding config.yaml files. They set a good example of how to use this package properly.

Procedure

DARLIN barcode induction

1. Induce DARLIN barcodes. Follow option A to induce barcoding at embryonic stages, or follow option B for neonatal or adult stage induction. This depends on the specific biological questions to be addressed; please refer to our DARLIN paper for specific examples[20].

   (A) **Barcode induction in embryonic mice**

   ● **TIMING** **12 h**

   (i) Set up timed pregnancy between homozygous Cas9-TdT-gRNAs-M2 mouse and homozygous CA/TA/RA triple target-array mouse in the evening based on the Jax protocol (https://www.jax.org/news-and-insights/jax-blog/2014/september/six-steps-for-setting-up-timed-pregnant-mice). Female Cas9-TdT-gRNAs-M2 mice and male CA/TA/RA mice are recommended to use to achieve high editing efficiency in embryos. The presence of a vaginal plug on the next morning indicates a successful pregnancy, with an embryonic stage corresponding to day 0.5 (E0.5).

   ◆ **TROUBLESHOOTING**

   (ii) To induce barcoding in developing embryos at the preferred stage, first measure the weight W of the pregnant dam on the injection day (unit: g), and then inject 50 × W μg Dox through a retro-orbital route of the dam.

   ▲ **CRITICAL STEP** Dox solution is freshly made.

   (B) **Barcode induction in neonatal or adult mice**

   ● **TIMING** **over 3 weeks**

   (i) Set up the cross between homozygous Cas9-TdT-gRNAs-M2 mouse and homozygous CA/TA/RA triple target-array mouse. The offspring of this cross are DARLIN mice.

(ii) To induce barcoding in neonatal or adult mice, measure the weight *W* of the DARLIN mouse and apply three intraperitoneal injections of 50 × *W* µg Dox every other day. During the same week, 2 mg/ml Dox water (supplemented with 10 mg/ml sucrose) is provided for the breast-feeding mothers (for neonatal induction) or the adult DARLIN mouse itself.

▲ **CRITICAL STEP** Dox solution is freshly made.

Evaluation of barcode induction efficiency

● **TIMING 1.5 h**

▲ **CRITICAL** To evaluate the barcode induction efficiency in DARLIN system, we use Phire Tissue Direct PCR Master Mix to quickly amplify three DARLIN arrays (CA, TA and RA) from tissues. This can be performed earlier using mouse tail or using the desired tissue on the sample collection day to help select the best edited embryo from a litter before cell sorting or a single cell RNA-sequencing experiment.

2. Put the tip of the mouse tail or a small amount of desired tissue into 20 µl of dilution buffer. Add 0.5 µl of DNARelease Additive. Vortex the tube briefly, and spin down the solution. Incubate the mixture at room temperature (~23 °C) for 5 min and place into 98 °C block for 2 min. 1 µl of the reaction mixture is used as template for DARLIN array amplification.
3. Prepare 20 µl of PCR reaction for CA, TA and RA, respectively, as follows:

CA PCR reaction	
Component	**Volume (20 µl rxn)**
2× Phire Tissue Direct PCR Master Mix	10 µl
CA_F primer (10 µM)	1 µl
CA_R primer (10 µM)	1 µl
Sample DNA template	1 µl
H_2O	7 µl
TA PCR reaction	
Component	**Volume (20 µl rxn)**
2× Phire Tissue Direct PCR Master Mix	10 µl
TA_F primer (10 µM)	1 µl
TA_R primer (10 µM)	1 µl
Sample DNA template	1 µl
H_2O	7 µl
RA PCR reaction	
Component	**Volume (20 µl rxn)**
2× Phire Tissue Direct PCR Master Mix	10 µl
RA_F primer (10 µM)	1 µl
RA_R primer (10 µM)	1 µl
Sample DNA template	1 µl
H_2O	7 µl

4. Run the PCR reaction with the following conditions:

Step	Temperature (°C)	Duration	Cycles
Initial denaturation	98	5 min	1
Denaturation	98	5 s	40
Annealing	64	5 s	
Extension	72	20 s	
Final extension	72	1 min	1
	4–10	Hold	1

5. Gel electrophoresis: load the PCR samples on a 2% (wt/vol) agarose gel, run and image the gel. Determine the barcode induction efficiency by comparing the intensity of the smear with the intensity of the control (unedited) band. Note that this is only a qualitative assay for quality control purpose. To actually quantify the editing efficiency, we would recommend using the sequencing approach described below.
   ◆ TROUBLESHOOTING

Bulk DARLIN library preparation and sequencing

▲ **CRITICAL** Follow Steps 6–23 for bulk DARLIN library preparation and sequencing. Alternatively, for single-cell applications follow Steps 24–37.

RNA extraction

● **TIMING** 2 h

6. RNA extraction can be performed on either tissue sample or FACS-sorted cells. It is advised to control the number of cells in a bulk sample before library preparation to avoid high sequencing cost.
   - Tissue lysis: place 1–10 mg tissue in a 1.5 ml tube, add 1 ml of TRIzol Reagent and homogenize the tissue with a homogenizer. The cell number can be estimated from the resulting RNA weight, since each cell has ~10–30 pg RNA.
   - Lysis of FACS-sorted cells: for 10^2–10^6 cells, add 1 ml of TRIzol Reagent, vortex thoroughly, centrifuge briefly and incubate at room temperature for 5 min. The cell number can be determined from FACS or with a cell counter.
     ■ **PAUSE POINT** Samples can be stored at −80 °C for up to a year.
7. Follow the User Guide of TRIzol Reagent (https://assets.thermofisher.com/TFS-Assets%2FLSG%2Fmanuals%2Ftrizol_reagent.pdf) for sample lysis, phase separation and RNA isolation (just follow this reagent user guide from Step 4 to the end of RNA isolation).
   ▲ **CRITICAL STEP** We highly recommend using glycogen to coprecipitate with sample RNA.
8. Quantify RNA yield and purity with NanoDrop spectrophotometer.
   ▲ **CRITICAL STEP** For samples with only a few cells, skip this step.
   ◆ TROUBLESHOOTING

Bulk DARLIN library preparation

● **TIMING** 1 d

9. Set up RNA denaturing reaction with the following condition in a 0.2 ml PCR tube, incubate the tube at 65 °C for 5 min and cool on ice for 1 min.

Component	Volume (13 µl reaction)
Sample RNA (500 ng or all sample if less than this weight)	9 µl
RT_CA_12UMI primer (10 µM)	1 µl
RT_TA_14UMI primer (10 µM)	1 µl
RT_RA_14UMI primer (10 µM)	1 µl
dNTPs (10 mM)	1 µl

▲ **CRITICAL STEP** Include a water-only control (that does not contain RNA) alongside samples. This will serve as a flag for sample cross-contamination and contamination of reagents with DARLIN DNA.

10. Add the following mixture to the RNA denature product, incubate the tube at 55 °C for 1 h and deactivate at 70 °C for 15 min.

Component	Volume (20 µl reaction)
RNA denature mixture	13 µl
5× First Strand Buffer	4 µl
DTT (0.1 M)	1 µl
RNAseOUT inhibitor	1 µl
Superscript III RT	1 µl

11. Purify the cDNA product with 1.5× AMPure XP beads (30 µl) once according to the manufacturer's protocol, and elute cDNA in 16.5 µl H_2O.

▲ **CRITICAL STEP** To efficiently capture short DARLIN arrays after editing, 1.5× AMPure XP beads are needed.

12. Add primers and reagents needed for the first round of DARLIN array amplification as follows:

Component	Volume (25 µl reaction)
5× Q5 DNA polymerase buffer	5 µl
dNTPs (10 mM)	0.5 µl
NGS_F primer (10 µM)	1.25 µl
NGS_CA_R1 primer (10 µM)	0.5 µl
NGS_TA_R1 primer (10 µM)	0.5 µl
NGS_RA_R1 primer (10 µM)	0.5 µl
cDNA product	16.5 µl
Q5 DNA polymerase	0.25 µl

13. Run the PCR reaction with the following conditions:

Step	Temperature (°C)	Duration	Cycles
Initial denaturation	98	30 s	1
Denaturation	98	10 s	12
Annealing	68	30 s	
Extension	72	30 s	
Final extension	72	2 min	1
	4–10	Hold	1

14. Purify the PCR product with 1.5× AMPure XP beads (37.5 µl) once according to the manufacturer's protocol, and elute the PCR product in 66 µl H_2O.

▲ **CRITICAL STEP** To efficiently capture short DARLIN arrays after editing, 1.5× AMPure XP beads are needed.

■ **PAUSE POINT** Samples can be stored at −20 °C for up to a year.

15. Setup the nested PCR reaction for CA, TA and RA separately, each with just one-quarter of the PCR product from the previous step.

CA nested PCR reaction	
Component	**Volume (25 µl reaction)**
5× Q5 DNA polymerase buffer	5 µl
dNTPs (10 mM)	0.5 µl
NGS_F primer (10 µM)	1.25 µl
NGS_CA_R2 primer (10 µM)	1.25 µl
cDNA product	16.5 µl
Q5 DNA polymerase	0.25 µl
TA nested PCR reaction	
Component	**Volume (25 µl reaction)**
5× Q5 DNA polymerase buffer	5 µl
dNTPs (10 mM)	0.5 µl
NGS_F primer (10 µM)	1.25 µl
NGS_TA_R2 primer (10 µM)	1.25 µl
cDNA product	16.5 µl
Q5 DNA polymerase	0.25 µl

RA nested PCR reaction	
Component	**Volume (25 µl reaction)**
5× Q5 DNA polymerase buffer	5 µl
dNTPs (10 mM)	0.5 µl
NGS_F primer (10 µM)	1.25 µl
NGS_RA_R2 primer (10 µM)	1.25 µl
cDNA product	16.5 µl
Q5 DNA polymerase	0.25 µl

▲ **CRITICAL STEP** Owing to the difference in RNA expression level among the tree DARLIN arrays, we perform PCR reaction for CA, TA and RA separately in this step.

16. Run the nested amplification reaction with the following conditions:

Step	Temperature (°C)	Duration	Cycles
Initial denaturation	98	30 s	1
Denaturation	98	10 s	12
Annealing	69	30 s	
Extension	72	30 s	
Final extension	72	2 min	1
	4–10	Hold	1

17. Purify the PCR product with 1.5× AMPure XP beads (37.5 µl) once according to the manufacturer's protocol, and elute the PCR product in 30 µl H_2O.
▲ **CRITICAL STEP** To efficiently capture short DARLIN arrays after editing, 1.5× AMPure XP beads are needed.
■ **PAUSE POINT** Samples can be stored at −20 °C for up to a year.

18. Setup the indexing PCR reaction for CA, TA and RA separately, each with 16.5 µl of the nested PCR product from the previous step.

Component	Volume (25 µl reaction)
5× Q5 DNA polymerase buffer	5 µl
dNTPs (10 mM)	0.5 µl
Index primer (10 µM)	2.5 µl
Nested PCR product	16.5 µl
Q5 DNA polymerase	0.5 µl

19. Run the indexing PCR reaction with the following conditions:

Step	Temperature (°C)	Duration	Cycles
Initial denaturation	98	30 s	1
Denaturation	98	10 s	10
Annealing	62	30 s	
Extension	72	30 s	
Final extension	72	2 min	1
	4–10	Hold	1

20. Purify the PCR product with 0.8× AMPure XP beads (20 µl) twice according to the manufacturer's protocol, and elute the PCR product in 20 µl H_2O.
▲ **CRITICAL STEP** Remaining primer dimers would influence the sequencing quality. To eliminate such effects, it is necessary to purify the DARLIN library twice with 0.8× AMPure XP beads.

21. Measure the concentration of DARLIN library on Qubit with dsDNA HS assay kit, run the DARLIN library on a Tape Station with High Sensitivity D5000 ScreenTape assay kit.
■ **PAUSE POINT** Samples can be stored at −20 °C for up to a year.

Sequence bulk DARLIN library

● **TIMING 2 d**

22. Pool individual DARLIN libraries properly by accounting for each library's cell number, expected read number and index sequence in particular. Quantify the pooled library with KAPA Library Quantification kit. For more information of library pooling, please refer to this online material: https://kb.10xgenomics.com/hc/en-us/articles/4440054574477-How-do-I-pool-10x-libraries-for-Illumina-sequencing.
23. Sequence the bulk DARLIN library on an Illumina MiSeq using a paired-end 500-cycle v2 kit (read 1: 250 cycles; i7 Index: 8 cycles; read 2: 250 cycles; Illumina, MS-102-2003) with 5% PhiX sequencing control v3 (Illumina, FC-110-3001) at a concentration of 10 pM. We recommend a sequencing depth of minimum 25 reads/cell for CA libraries, 50 reads/cell for TA libraries and 25 reads/cell for RA libraries. As mentioned earlier, we recommend limiting the number of cells in a bulk sample before library preparation to reduce unnecessary cost.
◆ **TROUBLESHOOTING**

DARLIN library preparation and sequencing from scRNA-seq libraries

▲ **CRITICAL** Follow Steps 24–37 for single-cell DARLIN library preparation and sequencing. Alternatively, for bulk applications follow Steps 6–23.

Amplification of DARLIN from scRNA-seq libraries

● **TIMING 1 d**

24. Follow the user guide of the Next-GEM Single-Cell 3′ Gene Expression v3.1 (user guide CG000315) for single-cell preparation, encapsulation, reverse transcription and initial cDNA amplification. At step 2.4 of the user protocol, after confirming the cDNA passes the quality control (QC), load 5 µl of cDNA into a PCR amplification reaction as follows:

CA PCR reaction	
Component	**Volume (25 µl reaction)**
Kapa HiFi HotStart ReadyMix	12.5 µl
P5_PR1 primer (20 µM)	0.75 µl
NGS_CA_R1 primer (10 µM)	0.75 µl
10× cDNA	5 µl
H_2O	6 µl
TA PCR reaction	
Component	**Volume (25 µl reaction)**
Kapa HiFi HotStart ReadyMix	12.5 µl
P5_PR1 primer (20 µM)	0.75 µl
NGS_TA_R1 primer (10 µM)	0.75 µl
10× cDNA	5 µl
H_2O	6 µl
RA PCR reaction	
Component	**Volume (25 µl reaction)**
Kapa HiFi HotStart ReadyMix	12.5 µl
P5_PR1 primer (20 µM)	0.75 µl
NGS_RA_R1 primer (10 µM)	0.75 µl
10× cDNA	5 µl
H_2O	6 µl

▲ **CRITICAL STEP** Ensure a 20 µM concentration of the P5_PR1 primer and 10µM concentration of the barcode array primers are used.

▲ **CRITICAL STEP** Ensure a water-only control reaction is setup alongside experimental reactions to allow detection of library cross-contamination.

25. Run the PCR reaction using the following program (lid temperature 105 °C):

Step	Temperature (°C)	Duration	Cycles
Initial denaturation	95	3 min	1
Denaturation	98	20 s	10
Annealing	65	15 s	
Extension	72	15 s	
Final extension	72	1 min	1
	4	Hold	1

26. Purify the cDNA product with 1.5× SPRIselect beads once according to the manufacturer's protocol, and elute cDNA in 20 µl Buffer EB.

▲ **CRITICAL STEP** To efficiently capture short DARLIN arrays after editing, 1.5× SPRIselect beads are needed.

27. Load 11 µl of library into a PCR amplification reaction as follows:

CA nested PCR reaction	
Component	**Volume (25 µl reaction)**
Kapa HiFi HotStart ReadyMix	12.5 µl
P5_PR1 primer (10 µM)	0.75 µl
NGS_CA_R2 primer (10 µM)	0.75 µl
CA library	11 µl
TA nested PCR reaction	
Component	**Volume (25 µl reaction)**
Kapa HiFi HotStart ReadyMix	12.5 µl
P5_PR1 primer (10 µM)	0.75 µl
NGS_TA_R2 primer (10 µM)	0.75 µl
TA library	11 µl
RA nested PCR reaction	
Component	**Volume (25 µl reaction)**
Kapa HiFi HotStart ReadyMix	12.5 µl
P5_PR1 primer (10 µM)	0.75 µl
NGS_RA_R2 primer (10 µM)	0.75 µl
RA library	11 µl

28. Run the PCR reaction using the following program (lid temperature 105 °C):

Step	Temperature (°C)	Duration	Cycles
Initial denaturation	95	3 min	1
Denaturation	98	20 s	10
Annealing	65	15 s	
Extension	72	15 s	
Final extension	72	1 min	1
	4	Hold	1

29. Purify the cDNA product with 1.5× SPRIselect beads once according to the manufacturer's protocol, and elute cDNA in 20 µl Buffer EB.

▲ **CRITICAL STEP** To efficiently capture short DARLIN arrays after editing, 1.5× SPRIselect beads are needed.

■ **PAUSE POINT** Samples can be stored at −20 °C for up to a year.

30. Load 11 µl of library into a PCR amplification reaction as follows:

Indexing PCR reaction	
Component	**Volume (25 µl reaction)**
Kapa HiFi HotStart ReadyMix	12.5 µl
Index from Chromium Dual Index kit	1.5 µl
Library from Step 29	11 µl

▲ **CRITICAL STEP** Ensure libraries that will be pooled have a unique index.

31. Run the PCR reaction using the following program (lid temperature 105 °C):

Step	Temperature (°C)	Duration	Cycles
Initial denaturation	95	3 min	1
Denaturation	98	20 s	9
Annealing	55	15 s	
Extension	72	15 s	
Final extension	72	1 min	1
	4	Hold	1

32. Measure the concentration of DARLIN library on a Qubit with the dsDNA HS assay kit and pool libraries by accounting for each library's cell number, expected read number and index sequence.
33. Purify the pooled PCR library with 0.8× SPRIselect beads twice according to the manufacturer's protocol and elute the PCR product in 20 µl Buffer EB.
    ■ **PAUSE POINT** Samples can be stored at −20 °C for up to a year.
34. Confirm amplification of libraries by TapeStation analysis or running on a 2% DNA agarose gel. Libraries should resemble a smear between 100 and 600 bp; a stronger band indicating the unedited barcode may be present at ~600 bp, depending on the editing efficiency of sample.

Sequence single-cell DARLIN library

● **TIMING 2 d**

35. Quantify the pooled library with the KAPA Library Quantification kit.
36. Calculate sequencing depth required. We recommend allocating 100× sequencing reads/cell for CA libraries, 200× sequencing reads/cell for TA libraries and 100× sequencing reads/cell for RA libraries.
37. Sequence the single-cell DARLIN library on an Illumina MiSeq using a paired-end 500-cycle v2 kit (read 1: 28 cycles; i7 Index: 8 cycles; read 2: 350 cycles; Illumina, MS-102-2003) with 5% PhiX sequencing control v3 (Illumina, FC-110-3001) at a concentration of 10 pM.
    ▲ **CRITICAL STEP** Ensure cycle times on read 1 and i7 matches the 10X protocol guidelines for the kit and indexes used.
    ◆ **TROUBLESHOOTING**

Data processing

● **TIMING 1 d**

38. Setup the project folder. To run the pipeline on a new dataset, create a new project folder. Also create a config.yaml file as well as a raw_fastq folder under this project folder. The config.yaml file can be copied from the template associated with the test data in snakemake_DARLIN package. The fastq files should be placed under the raw_fastq folder, and match the naming convention $SAMPLE_R{1,2}.fastq.gz.
39. Update the config file. Modify parameters in the pipeline config file using a text editor. Change the list of selected samples, library type (that is, bulk or single-cell libraries), DARLIN reference template, read cutoff for calling call barcodes and unique molecular identifiers (UMIs) and so on, according to your setup. See Table 4 for parameter description.

Table 4 | key parameters in the config.yaml file for running the snakemake_DARLIN pipeline

Parameter name	Explanation	Example
SampleList	A list of sample names to be processed, matching the names in the raw_fastq files: $SAMPLE_L001_R{1,2}_001.fastq.gz	['sample1', 'sample2', 'sample_3']
cfg_type	Library type BulkRNA_12UMI: bulk DARLIN library from CA, with 12 bp UMI BulkRNA_Tigre_14UMI: bulk DARLIN library from TA, with 14 bp UMI BulkRNA_Rosa_14UMI: bulk DARLIN library from RA, with 14 bp UMI sc10xV3: single-cell DARLIN library from 10X scCamellia: single-cell DARLIN library from Camellia-seq	'BulkRNA_12UMI'
template	Unedited DARLIN sequence template cCARLIN: template for CA library Tigre_2022_v2: template for TA library Rosa_v2: template for RA library	'cCARLIN'
read_cutoff_UMI_override	A list of all UMI cutoffs to be used. Results for each cutoff will be saved to a separate folder. We recommend a minimum cutoff of 3	[3, 10]
sbatch	Whether to submit the job to SLURM. 1: submit jobs; 0: run the computation locally	1

Note that samples included in one config file should have the same library type, template and read cutoff. If there are multiple samples of different type, please separate them into different project folders or use a separate config file.

40. Run the following command under the project folder where the config.yaml file is located, to process each of the samples specified in the config file. This will generate a result folder separately for each sample.

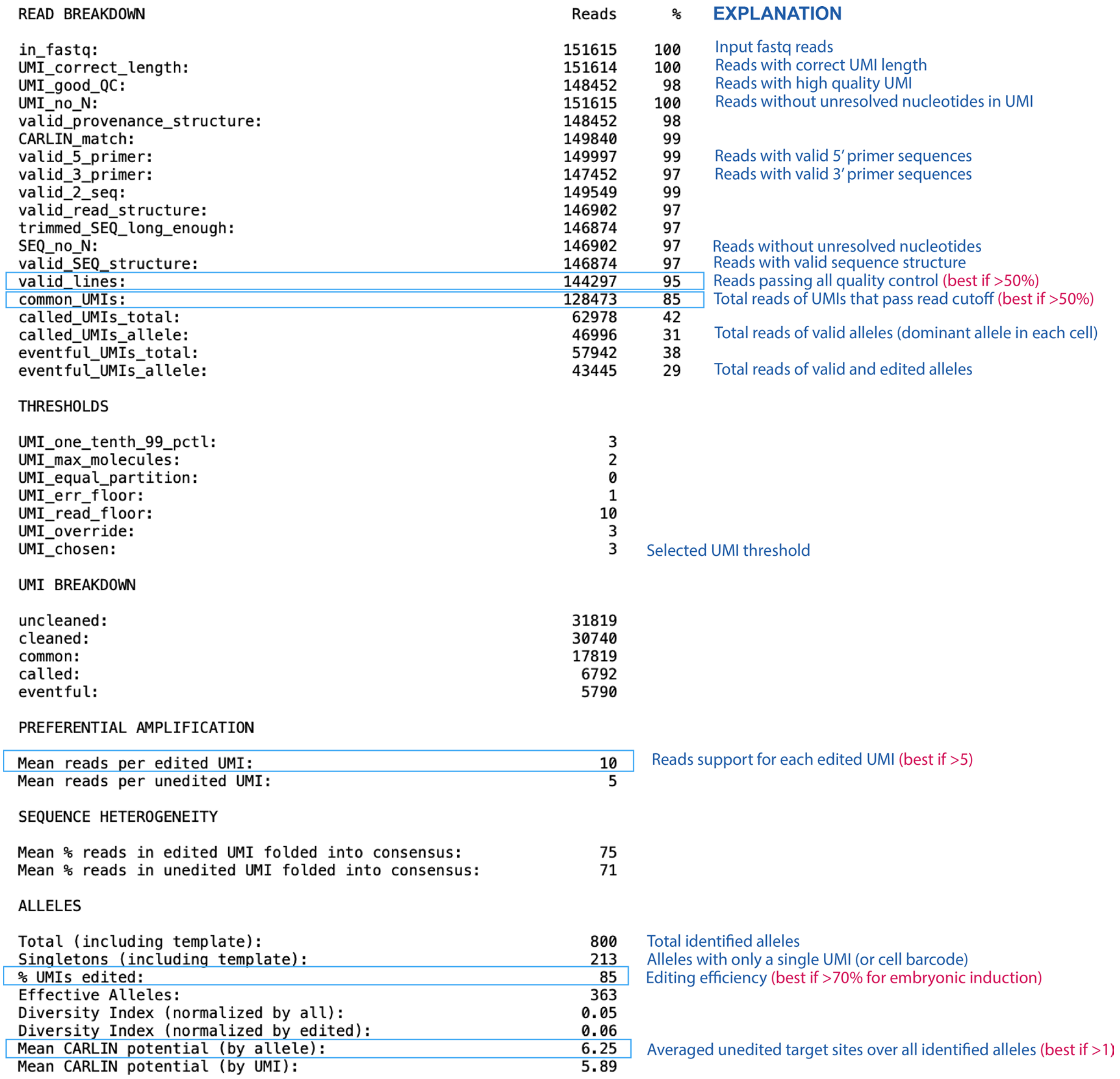
```
snakemake -s $snakemake_DARLIN_path/snakefiles/snakefile_matlab_
DARLIN_Part1.py --configfile config.yaml --core 10 --ri
```

41. Run the second command at the same location to generate an aggregated report across multiple samples. The result will be saved at a merge_all folder.

```
snakemake -s $snakemake_DARLIN_path/snakefiles/snakefile_matlab_
DARLIN_Part2.py --configfile config.yaml --core 5 --ri
```

42. Inspect key intermediate output files from the above preprocessing, which are located at `PROJECT/DARLIN/results_cutoff_override_$CUTOFF`, including:
    - `$SAMPLE/Results.txt`: summary statistics of a sample, including read number breakdown at each QC step, the mean reads per edited cell barcode or UMI, the total allele number and so on. See Fig. 8 for a representative output and the expected results
    - `$SAMPLE/AlleleAnnotation.txt`: alleles, with mutation pattern encoded in text strings, detected in a given sample
    - `$SAMPLE/AlleleColonies.txt`: corresponding cell barcode or UMI information of each allele observed in AlleleAnnotation.txt
    - `$SAMPLE/allele_UMI_count.csv`: a table of the UMI (or cell barcode) counts for each allele in a given sample
    - `merge_all/allele_UMI_count.csv`: a table of the UMI (or cell barcode) counts for each allele across all samples
    - `merge_all/refined_results.csv`: a table of summary statistics across all samples
    - `merge_all/DARLIN_report.html`: a html report including key QC figures and tables across all samples

    ◆ **TROUBLESHOOTING**



```
READ BREAKDOWN                                        Reads     %    EXPLANATION

in_fastq:                                            151615   100    Input fastq reads
UMI_correct_length:                                  151614   100    Reads with correct UMI length
UMI_good_QC:                                         148452    98    Reads with high quality UMI
UMI_no_N:                                            151615   100    Reads without unresolved nucleotides in UMI
valid_provenance_structure:                          148452    98
CARLIN_match:                                        149840    99
valid_5_primer:                                      149997    99    Reads with valid 5′ primer sequences
valid_3_primer:                                      147452    97    Reads with valid 3′ primer sequences
valid_2_seq:                                         149549    99
valid_read_structure:                                146902    97
trimmed_SEQ_long_enough:                             146874    97
SEQ_no_N:                                            146902    97    Reads without unresolved nucleotides
valid_SEQ_structure:                                 146874    97    Reads with valid sequence structure
valid_lines:                                         144297    95    Reads passing all quality control (best if >50%)
common_UMIs:                                         128473    85    Total reads of UMIs that pass read cutoff (best if >50%)
called_UMIs_total:                                    62978    42
called_UMIs_allele:                                   46996    31    Total reads of valid alleles (dominant allele in each cell)
eventful_UMIs_total:                                  57942    38
eventful_UMIs_allele:                                 43445    29    Total reads of valid and edited alleles

THRESHOLDS

UMI_one_tenth_99_pctl:                                    3
UMI_max_molecules:                                        2
UMI_equal_partition:                                      0
UMI_err_floor:                                            1
UMI_read_floor:                                          10
UMI_override:                                             3
UMI_chosen:                                               3   Selected UMI threshold

UMI BREAKDOWN

uncleaned:                                            31819
cleaned:                                              30740
common:                                               17819
called:                                                6792
eventful:                                              5790

PREFERENTIAL AMPLIFICATION

Mean reads per edited UMI:                               10   Reads support for each edited UMI (best if >5)
Mean reads per unedited UMI:                              5

SEQUENCE HETEROGENEITY

Mean % reads in edited UMI folded into consensus:        75
Mean % reads in unedited UMI folded into consensus:      71

ALLELES

Total (including template):                             800   Total identified alleles
Singletons (including template):                        213   Alleles with only a single UMI (or cell barcode)
% UMIs edited:                                           85   Editing efficiency (best if >70% for embryonic induction)
Effective Alleles:                                      363
Diversity Index (normalized by all):                   0.05
Diversity Index (normalized by edited):                0.06
Mean CARLIN potential (by allele):                     6.25   Averaged unedited target sites over all identified alleles (best if >1)
Mean CARLIN potential (by UMI):                        5.89
```

Fig. 8 | A representative QC report from the Results.txt file after data processing. Output file from Step 42. Key metrics are highlighted with a blue box.

43. For downstream DARLIN data analysis on these intermediate output files, please follow our step-by-step tutorial (https://github.com/ShouWenWang-Lab/DARLIN_tutorial), which includes common allele filtering, clone identification, shared clone fraction analysis, lineage-coupling analysis between tissues or cell types and integration with single-cell transcriptome to infer cell-fate bias.

Troubleshooting

Troubleshooting advice for the experimental procedure can be found in Table 5 and for data processing in Table 6. For data processing, see also Fig. 8.

Table 5 | Troubleshooting table for experimental procedures

Step	Problem	Possible reason	Solution
1A(i)	Failed timed pregnancy	The female mouse is not in proestrus or estrus	Setting up timed pregnancy for mice at estrous state can increase the likelihood of pregnancy
5	Low barcode induction efficiency in adults	The editing efficiency in adults varies among individuals	Prepare more induced adults and choose the ones with highest editing for DARLIN array analysis
8	The A260/280 ratio is low in RNA extraction	The organic phase is not removed completely	Do not pipette up the entire aqueous layer after phase separation
23	High-cluster density in sequencing	The complexity of the DARLIN library is low	Reduce the library loading concentration and add more PhiX
37	Poor sequencing read quality (reads PF <80% or Q30 <60%)	Overloaded sequencing. Lower loading concentrations may be required for single-cell DARLIN sequencing due to the long cycle number on read 2	Reduce library loading concentration
23,37	DNA present in water-only control	Cross-contamination of samples or contamination of one or more reagents with DARLIN DNA	Clean bench and pipettes with 10% bleach and/or use designated 'clean' library-preparation workspace. Ensure extreme care is taken to avoid cross-contamination of samples and reagents

Table 6 | Troubleshooting table for data processing

Step	Problem	Possible reason	Solution
42	Results.txt: Fraction of valid_lines is low	Low sequencing quality	Resequence the data. Also try to increase library complexity
	Results.txt: Fraction of common_UMIs is low	Low sequencing depth or UMI read cutoff too high	Increase sequencing depth or decrease the UMI read cutoff
	Results.txt: Mean reads per edited UMI is low (e.g., ≤3)	Low sequencing depth	Increase sequencing depth
	Results.txt: % UMIs edited is low	Insufficient Dox induction	Improve the Dox induction procedure
	Results.txt: Mean CARLIN potential by allele is close to 0	Maybe due to exhaustive editing and may reduce allele diversity and increase barcode homoplasy	This could lead to over editing. Try to reduce the Dox concentration or induction duration

Timing

Step 1, timed pregnancy: 12 h, including hands-on time of 30 min
Step 1, Dox induction: 30 min
Steps 2–5, estimation of barcode induction efficiency: 1.5 h
Steps 6–8, RNA extraction: 2 h
Steps 9–20, bulk DARLIN library preparation, 1 d
Step 21, QC of the bulk library, 2 h
Step 22, bulk library pooling and quantification, 2 h
Step 23, bulk library sequencing, 2 d, 2 h hands-on time
Steps 24–34, single-cell library preparation, 1 d
Steps 35,36, single-cell library pooling and quantification, 2 h
Step 37, single-cell library sequencing, 2 d, 2 h hands-on time
Steps 38–43, data processing, ~1 d

Anticipated results

Before sequencing, the editing efficiency can be estimated by amplifying CA, TA and RA, respectively, from the genomic DNA extracted from the mouse tail or desired tissues, and running an agarose gel analysis or TapeStation (Figs. 4 and 5). One may select the mice with higher editing according to this editing analysis to proceed with bulk or single-cell sequencing of DARLIN arrays. For induction at embryonic stage, ~90% editing efficiency is expected if DARLIN mice are generated with female Cas9-TdT-gRNAs-M2 mice and male CA/TA/RA mice. For lineage labeling at the adult stage, the editing efficiency can be more variable (Table 2).

In single-cell sequencing, users can expect to recover barcodes (edited and unedited) from ~30% (CA), ~60% (TA) and ~35% (RA) of total cells in which whole transcriptome information is available. Together, ~80% of cells have detected barcodes from at least one locus. For just edited barcodes, we observed a recovery efficiency of ~25%, ~55% and ~25% in single cells from CA, TA and RA, respectively, or ~70% from at least one locus. After removing common barcodes that suffer from barcode homoplasy, one can integrate the barcode information across different loci to identify clones (Figs. 6 and 7).

The number of detected clones depends on the induction timepoint, specific cell populations that are analyzed, the number of cells used for analysis, sequencing depth and the read cutoff used for QC[20]. Induction at the embryonic stage typically results in fewer clones due to fewer cells that are initially labeled, while labeling cells at adult stage may result in many more clones.

Once reliable clones are identified following our analysis guidelines, whether from single-cell or bulk sequencing data, one may carry out further downstream analyses, which are probably specific to each biological system or problem. We summarized useful clonal analyses in Fig. 6b–e, which were performed in our recent DARLIN paper[20]. These include clone visualization on transcriptomic embedding and fate bias prediction from CoSpar among early progenitors. These analyses are applicable to only single-cell datasets. For both bulk and single-cell data, one can calculate the shared clone fraction between different cell populations, which quantifies the fraction of clones that are jointly detected in two populations, and the clonal coupling score between different cell populations, which evaluates how much two cell populations share similar developmental origins. Individual clones typically show localized structure in the transcriptomic embedding, reflecting certain clonal fate bias, although unbiased and multipotent clones also exist (Fig. 6b). In physical space, clones are also expected to be local soon after barcode induction, before migrating to other locations later, which are evident from our data collected from different bones in the same mice (Fig. 6c,d). We found that myeloid progenitors migrate/circulate much faster to other bones compared with hematopoietic stem cells (Fig. 6c). In the blood system, CoSpar prediction based on the DARLIN data revealed a HSC subtype that shows distinct transcriptomic feature and has high probability to generate megakaryocyte (Fig. 6e). We have developed a step-by-step tutorial on all the four analyses using our published DARLIN datasets, which is available in https://github.com/ShouWenWang-Lab/DARLIN_tutorial.

Data availability

Raw and intermediate data associated with this tutorial can be obtained in https://zenodo.org/records/11929508.

Code availability

The snakemake_DARLIN package for DARLIN data processing is available at https://github.com/ShouWenWang-Lab/snakemake_DARLIN. The companion Python package for downstream analysis is available at https://github.com/ShouWenWang-Lab/MosaicLineage. A tutorial for downstream analyses written in jupyter notebooks can be found at https://github.com/ShouWenWang-Lab/DARLIN_tutorial.

Received: 27 June 2024; Accepted: 7 January 2025;
Published online: 21 March 2025

References

1. Kretzschmar, K. & Watt, F. M. Lineage tracing. *Cell* **148**, 33–45 (2012).
2. Bałakier, H. & Pedersen, R. A. Allocation of cells to inner cell mass and trophectoderm lineages in preimplantation mouse embryos. *Dev. Biol*. **90**, 352–362 (1982).
3. Snippert, H. J. et al. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).
4. He, L. et al. Enhancing the precision of genetic lineage tracing using dual recombinases. *Nat. Med*. **23**, 1488–1498 (2017).
5. Liu, K. et al. Tracing the origin of alveolar stem cells in lung repair and regeneration. *Cell* **187**, 2428–2445.e20 (2024).
6. He, L. et al. Proliferation tracing reveals regional hepatocyte generation in liver homeostasis and repair. *Science* **371**, eabc4346 (2021).
7. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
8. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol*. **36**, 442–450 (2018).
9. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol*. **36**, 469–473 (2018).
10. 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).
11. Chan, M. M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).
12. Kalhor, R. et al. Developmental barcoding of whole mouse via homing CRISPR. *Science* **361**, eaat9804 (2018).
13. Pei, W. et al. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* **548**, 456–460 (2017).
14. Sun, J. et al. Clonal dynamics of native haematopoiesis. *Nature* **514**, 322–327 (2014).
15. Pei, W. et al. Resolving fates and single-cell transcriptomes of hematopoietic stem cell clones by PolyloxExpress barcoding. *Cell Stem Cell* **27**, 383–395.e8 (2020).
16. Weber, T. S. et al. LoxCode in vivo barcoding resolves epiblast clonal fate to fetal organs. Preprint at *bioRxiv* https://doi.org/10.1101/2023.01.02.522501 (2023).
17. Xie, L. et al. Comprehensive spatiotemporal mapping of single-cell lineages in developing mouse brain by CRISPR-based barcoding. *Nat. Methods* https://doi.org/10.1038/s41592-023-01947-3 (2023).
18. Liu, K. et al. Mapping single-cell-resolution cell phylogeny reveals cell population dynamics during organ development. *Nat. Methods* https://doi.org/10.1038/s41592-021-01325-x (2021).
19. 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**, 1410–1422.e27 (2020).
20. Li, L. et al. A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells. *Cell* https://doi.org/10.1016/j.cell.2023.09.019 (2023).
21. Wang, S.-W., Herriges, M. J., Hurley, K., Kotton, D. N. & Klein, A. M. CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nat. Biotechnol*. https://doi.org/10.1038/s41587-022-01209-1 (2022).
22. Wang, K. et al. PhyloVelo enhances transcriptomic velocity field mapping using monotonically expressed genes. *Nat. Biotechnol*. https://doi.org/10.1038/s41587-023-01887-5 (2023).
23. Jones, M. G. et al. Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. *Genome Biol*. **21**, 92 (2020).
24. Schiffman, J. S. et al. Defining heritability, plasticity, and transition dynamics of cellular phenotypes in somatic evolution. *Nat. Genet*. https://doi.org/10.1038/s41588-024-01920-6 (2024).
25. Patel, S. H. et al. Lifelong multilineage contribution by embryonic-born blood progenitors. *Nature* **606**, 747–753 (2022).
26. Abdullah, L. et al. Hierarchal single-cell lineage tracing reveals differential fate commitment of CD8 T-cell clones in response to acute infection. Preprint at *bioRxiv* https://doi.org/10.1101/2024.03.21.586160 (2024).
27. Rodriguez-Fraticelli, A. E. et al. Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. *Nature* https://doi.org/10.1038/s41586-020-2503-6 (2020).
28. Rodriguez-Fraticelli, A. E. et al. Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212–216 (2018).
29. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, eaaw3381 (2020).
30. Weng, C. et al. Deciphering cell states and genealogies of human hematopoiesis. *Nature* https://doi.org/10.1038/s41586-024-07066-z (2024).
31. Engblom, C. et al. Spatial transcriptomics of B cell and T cell receptors reveals lymphocyte clonal dynamics. *Science* **382**, eadf8486 (2023).
32. Chen, A. et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777–1792.e21 (2022).
33. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
34. Tian, L., Chen, F. & Macosko, E. Z. The expanding vistas of spatial transcriptomics. *Nat. Biotechnol*. **41**, 773–782 (2023).
35. Beard, C., Hochedlinger, K., Plath, K., Wutz, A. & Jaenisch, R. Efficient method to generate single-copy transgenic mice by site-specific integration in embryonic stem cells. *Genesis* **44**, 23–28 (2006).

Acknowledgements

The authors would like to thank F. Garcia Osorio for his work on developing the Cas9–CARLIN model and initial barcode amplification protocols. L.L. would like to thank H. Wu, a former research assistant in her laboratory, for plotting the schematic genetics of DARLIN mouse. S.-W.W. would like to thank M. Chen, a PhD student in his laboratory, for formatting this manuscript. S.-W.W. acknowledges support from Westlake High-Performance Computing Center. L.L. is supported by National Key Research and Development Project of China (2024YFA1306603). L.L. and S.-W.W. acknowledge support from 'Pioneer' and 'Leading Goose' R&D Programs of Zhejiang province (2024SSYS0034 and 2024SSYS0034). S.B. acknowledges support from the NIH (grant number 1K99HL164969). S.-W.W. acknowledges support from the NSFC (32470700). F.D.C. is funded by grants R01HL128850, RC2DK131963 and R01HL158192, and an Alex Lemonade Crazy 8 award.

Author contributions

This protocol is based on a paper by L.L., S.B., S.-W.W. and F.D.C., where L.L. and S.B. developed the DARLIN model with input from F.D.C., and S.-W.W. developed and carried out computational analyses. Here, S.-W.W., L.L. and S.B. wrote the manuscript. H.L. helped to develop the analyses tutorial and generated Figs. 3 and 7 with supervision from S.-W.W.; and D.C. generated Figs. 2, 4 and 5 with supervision from L.L.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Shou-Wen Wang or Fernando D. Camargo.

Peer review information *Nature Protocols* thanks Zheng Hu, Bushra Raj and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.