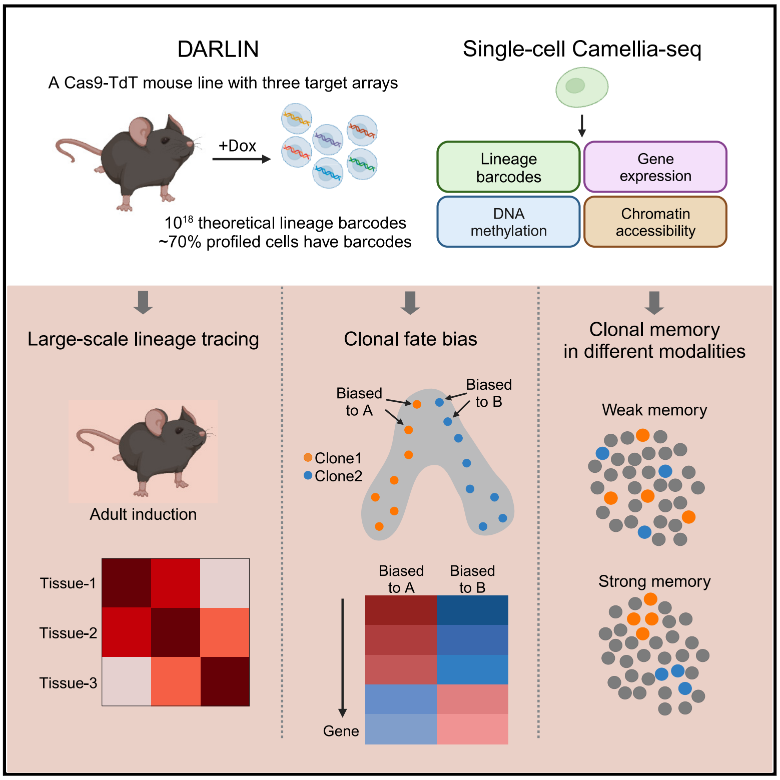
Cell

A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells

Graphical abstract



Authors

Li Li, Sarah Bowling, Sean E. McGeary, ..., Allon M. Klein, Shou-Wen Wang, Fernando D. Camargo

Correspondence

wangshouwen@westlake.edu.cn (S.-W.W.), fernando.camargo@childrens.harvard.edu (F.D.C.)

In brief

DARLIN is an inducible barcoding system that allows for lineage tracing and analysis across mouse tissues as well as combined transcriptional and epigenomic single-cell measurements.

Highlights

- DARLIN generates massive barcode diversity and labels ∼70% of profiled cells
- DARLIN identifies early fate bias among HSCs and their transcriptomic signatures
- DARLIN reveals low-level HSC circulation between bone-marrow niches in adulthood
- Strong clonal memory in DNA methylation rather than mRNA or chromatin accessibility

Li et al., 2023, Cell *186*, 1–17
November 9, 2023 © 2023 Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2023.09.019

CellPress

Cell

CellPress

Resource

A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells

Li Li,[1,2] Sarah Bowling,[1,2] Sean E. McGeary,[3] Qi Yu,[1,2] Bianca Lemke,[1,2] Karel Alcedo,[1,2] Yuemeng Jia,[1,2] Xugeng Liu,[1,2] Mark Ferreira,[1,2] Allon M. Klein,[3] Shou-Wen Wang,[4,5,6,*] and Fernando D. Camargo[1,2,7,*]
[1]Stem Cell Program, Boston Children's Hospital, Boston, MA, USA
[2]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA
[3]Department of Systems Biology, Blavatnik Institute, Harvard Medical School, Boston, MA 02115, USA
[4]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, China
[5]School of Life Sciences, Westlake University, Hangzhou, Zhejiang 310024, China
[6]School of Science, Westlake University, Hangzhou, Zhejiang 310024, China
[7]Lead contact
*Correspondence: wangshouwen@westlake.edu.cn (S.-W.W.), fernando.camargo@childrens.harvard.edu (F.D.C.)
https://doi.org/10.1016/j.cell.2023.09.019

SUMMARY

Cellular lineage histories and their molecular states encode fundamental principles of tissue development and homeostasis. Current lineage-recording mouse models have insufficient barcode diversity and single-cell lineage coverage for profiling tissues composed of millions of cells. Here, we developed DARLIN, an inducible Cas9 barcoding mouse line that utilizes terminal deoxynucleotidyl transferase (TdT) and 30 CRISPR target sites. DARLIN is inducible, generates massive lineage barcodes across tissues, and enables the detection of edited barcodes in ∼70% of profiled single cells. Using DARLIN, we examined fate bias within developing hematopoietic stem cells (HSCs) and revealed unique features of HSC migration. Additionally, we established a protocol for joint transcriptomic and epigenomic single-cell measurements with DARLIN and found that cellular clonal memory is associated with genome-wide DNA methylation rather than gene expression or chromatin accessibility. DARLIN will enable the high-resolution study of lineage relationships and their molecular signatures in diverse tissues and physiological contexts.

INTRODUCTION

Tracing cellular lineage history in animals has been a long-standing effort. Historically, labeling cells with distinguishable and heritable markers such as dyes has led to major discoveries in early development and stem cell differentiation.[1–3] However, this approach is limited to tracking only small or pre-defined populations of cells. Retrovirally barcoding cells with synthetic DNA sequences has enabled analysis of much larger populations,[4,5] although this requires *ex vivo* manipulation of cells. *In vivo* DNA barcoding in mouse models has been achieved through the use of randomly integrated transposons or recombinases that create genetic diversity within a distinct locus, which revealed a drastically different picture of hematopoiesis *in vivo*.[6–10] However, these mouse models either have limited barcode diversity or do not allow simultaneous interrogation of lineage and state information in single cells.

The advent of CRISPR-Cas9 technology has created a new avenue for lineage tracing where diverse DNA mutations can be created within a defined locus through genome editing.[11] The mutational outcomes can be transcribed, thereby allowing joint measurement of lineage and transcriptomic information in single cells.[12–14] In mice, these approaches have been used to study early embryonic development[15,16] and cancer progression.[17,18] Applying the same tools, we developed Cas9/CARLIN, a stable and genetically defined mouse line, which enables flexible induction at any point to generate diverse, transcribed lineage barcodes across tissues.[19]

These and other single-cell lineage-tracing approaches have generally faced three technical challenges: (1) low lineage-barcode capture efficiency in single-cell readout; (2) low efficiency of introducing lineage barcodes; and (3) contamination from barcode homoplasy, where an identical editing event occurs independently in two different cells. As a result of these challenges, only ∼10% of profiled cells from the Cas9/CARLIN mouse contain detected lineage barcodes that likely label individual clones.[19] Therefore, a higher-performing lineage-tracing mouse line is needed to enable high-coverage single-cell lineage tracing in adult tissues with millions of cells.

Single-cell lineage tracing with transcriptomic measurements has been successfully used to identify early fate bias among progenitors and find novel regulators of cell-fate choices.[20–22]

However, epigenomic modalities such as chromatin accessibility and DNA methylation are known to play a crucial role in regulating gene expression and maintaining cell identities.[23–25] Epigenetic changes are known to foreshadow changes in gene expression,[26–29] suggesting that the earliest events for cell-fate choice are unlikely to be captured using gene expression alone. This view is supported by a recent state-fate lineage-tracing study showing that the transcriptome of a cell alone is insufficient to predict its fate outcome.[21] To fully understand cell-fate choices and the maintenance of cell identities, single-cell approaches that integrate lineage, transcriptomic, and epigenomic information will be necessary. Several recent studies have reported single-cell measurement of either lineage and transcriptomic information, lineage and epigenomic information, or transcriptomic and epigenomic information.[25,30–32] Although some multi-omic studies have inferred lineage information using endogenous DNA mutations,[33,34] the inferred clones are low-resolution and cannot be used to study lineage relationships at defined developmental time points. Indeed, a method capable of simultaneously profiling engineered lineage barcodes, the transcriptome, and the epigenome in single cells has not been reported.

Here, we developed an improved lineage-tracing mouse line (DARLIN) that has an extremely large lineage-barcode capacity and highly efficient lineage recovery in single-cell assays, greatly outperforming the Cas9/CARLIN model. Furthermore, we extended existing approaches to simultaneously measure DNA methylation, chromatin accessibility, gene expression, and lineage information in single cells. We utilized DARLIN and its associated analysis tools to address three distinct lineage-related problems in hematopoiesis.

RESULTS

Cas9-TdT introduces more insertions than Cas9 upon transient induction in CARLIN mice

CRISPR-Cas9-based DNA editing is prone to deletions, which limits the resulting barcode diversity. Reanalyzing the editing events observed among the 10 tandem target sites within the integrated locus (referred to as target array) from the published Cas9/CARLIN mouse dataset[19] (Figure 1A), we found that deletions occurred more frequently than insertions (Figure 1B), with 1.5 insertion events and 2.5 deletion events per allele on average. An allele generated by Cas9 editing had a median of 163 bp deleted out of its 270-bp unedited target array, implying the deletion of 6 out of 10 tandem target sites (Figure 1C). By contrast, an allele has only a median of 2 bp inserted (Figure 1D). These large deletion events can lead to information loss and generate degenerate alleles. Consistent with this, common alleles were enriched with deletion-only alleles, whereas rare alleles, which are required for confident assignment of *bona fide* clones, preferentially resulted from DNA insertions (Figure 1E).

We reasoned that increasing the frequency of insertions could greatly increase the generation of rare alleles, thereby increasing overall allele diversity and barcoding capacity. Terminal deoxynucleotidyl transferase (TdT) is a template-independent DNA polymerase that can insert random nucleotides at both overhang and blunt 3′ ends.[35–37] A recent study in cell lines showed that co-expression of Cas9 and TdT generates more insertions in the target sites and leads to higher barcode diversity compared with Cas9 alone (Figure 1F).[38] To test whether a similar strategy also works in an organismal context, we hydrodynamically injected a plasmid encoding either a Cas9-TdT fusion protein or native Cas9 into the tail veins of adult mice carrying the target array (Figure 1A) and analyzed the resulting allele editing in mouse livers with bulk RNA sequencing (RNA-seq) after 1 week (Figure 1G). We observed that Cas9-TdT expression resulted in fewer deletions but twice the insertion events per allele than Cas9 expression (Figures 1H–1J). Aggregating insertion events from all alleles, we also observed more inserted nucleotides per allele upon Cas9-TdT expression (Figure 1K), with all four nucleotides well-represented in the inserted sequences (Figure 1L). These data demonstrate that Cas9-TdT introduces more insertions as well as fewer deletions than Cas9 upon target-array editing in an adult mouse.

DARLIN: An inducible Cas9-TdT mouse line for high-capacity lineage tracing

Having established that the Cas9-TdT expression leads to improved lineage-barcode editing in an adult mouse, we set out to generate an inducible germline mouse model that utilizes Cas9-TdT for target-array editing. First, we created a mouse embryonic stem cell (mESC) line with a Dox-inducible Cas9-TdT construct also carrying a CRISPR target array and cognate gRNAs (Figure S1A). We validated that editing of the array in this mESC line was sensitive to Dox exposure (Figures S1A–S1D).

We next engineered knockin mice with the tetO-Cas9-TdT construct inserted into the *Col1a1* locus.[39] This line was crossed with animals containing the gRNAs and *Col1a1* target array (CA) previously described in the Cas9/CARLIN system to generate *Col1a1*$^{\text{tetO-Cas9-TdT/gRNA-Array}}$:*Rosa26*$^{\text{M2-rtTA/+}}$ mice for lineage barcoding. We will refer to this particular line as DARLIN-v0 (Cas9-TdT CARLIN version 0; Figure 2A). To benchmark DARLIN-v0 against the original Cas9/CARLIN mouse line, we compared the allele editing observed in large numbers of granulocytes from each mouse line after 1 week of Dox treatment, followed by another 3 days without Dox (Figure 2B; n = 7 mouse replicates for DARLIN-v0 and n = 5 for Cas9/CARLIN). Compared with Cas9/CARLIN, edited alleles from the DARLIN-v0 mouse were enriched in rare alleles (Figures 2C and S1E). Indeed, for mouse replicates with over 10,000 alleles, ~65% of alleles were observed only once (referred to as singleton alleles) for DARLIN-v0, compared with ~30% for Cas9/CARLIN (Figure 2D). For the same number of edited cells (i.e., UMIs), the DARLIN-v0 mice exhibited 2.3-fold as many alleles as Cas9/CARLIN (Figure S1F). Since the utility of an allele for clonal labeling depends on its occurrence frequency, we used the metric 2^H, where H is the Shannon entropy of the normalized allele frequency across edited cells, to report the barcode diversity of an ensemble of alleles.[19] The Shannon allele diversity of DARLIN-v0 alleles was ~5 times that of Cas9/CARLIN alleles (Figure 2E). Consistent with this, we observed that DARLIN-v0 not only had much fewer large-scale (>180 bp) deletions that result in degenerate alleles (Figure 2F) but also more (Figure S1G) and larger (Figure 2G) insertions. Considering all mutation events across alleles, each insertion in Cas9/CARLIN was on average accompanied by three deletions, whereas in DARLIN-v0, each insertion was accompanied by fewer than one deletion

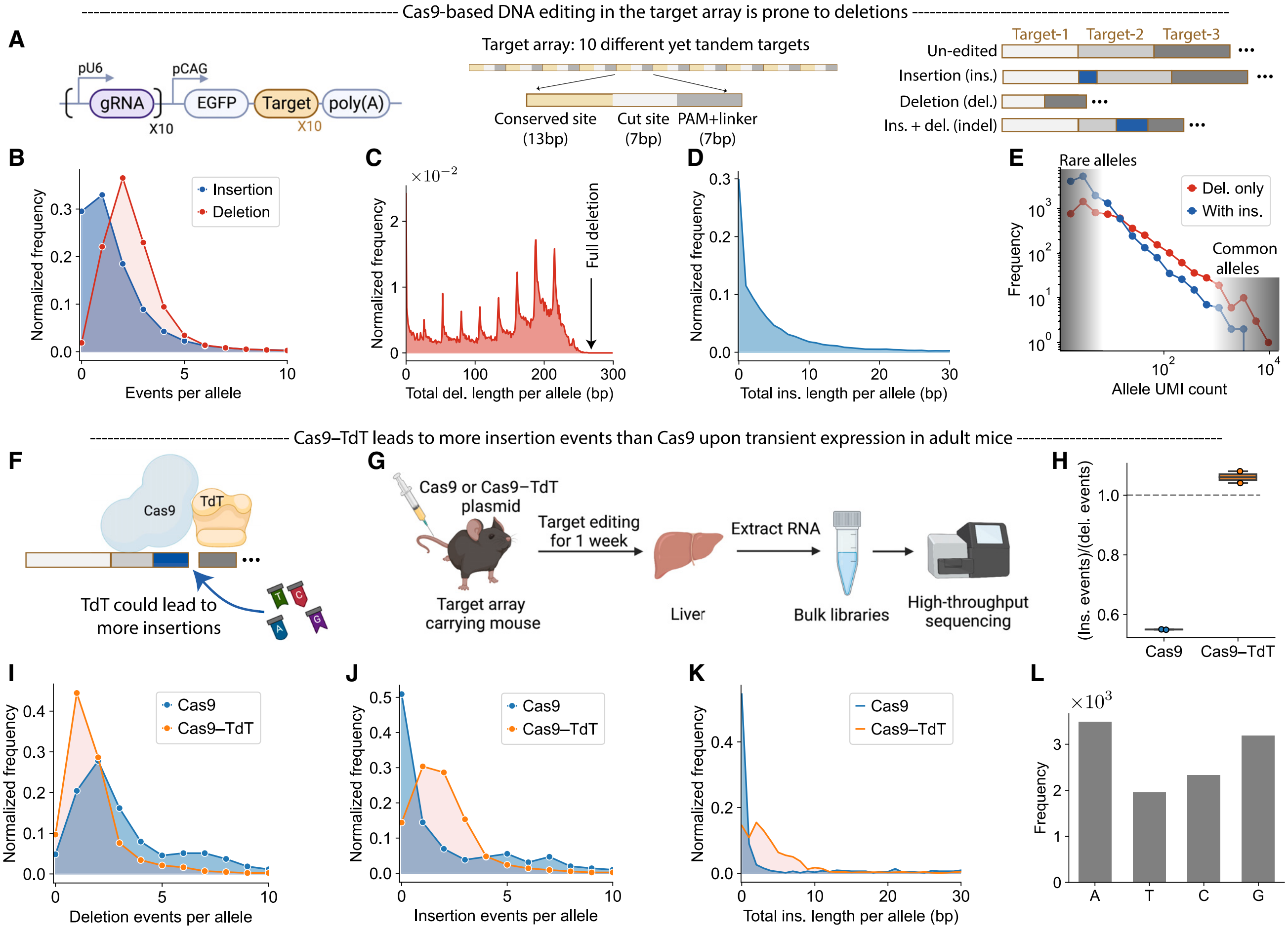


Figure 1. Advantage of Cas9-TdT over Cas9 for generating lineage barcodes

(A) Schematic of CARLIN lineage-recording system (left), the target array (middle), and editing patterns (right).

(B–E) Reanalysis of published bulk granulocyte Cas9/CARLIN data.[19]

(B–D) Distribution of insertion or deletion events (B), total deletion length (C), or total insertion length (D) per edited allele.

(E) Histogram of allele UMI counts among edited alleles, which either only contain deletions or have insertions and possibly deletions. UMI, unique molecular identifier.

(F) Target-array editing by a Cas9-TdT fusion protein.

(G) Experimental scheme to compare the target-array editing from Cas9 or Cas9-TdT protein.

(H) Box plot of the insertion-to-deletion event ratio among all edited alleles from Cas9 or Cas9-TdT mice. Horizontal lines of each box represent the minimum, 25th-, 50th-, 75th-percentiles, and maximum values.

(I–K) Distribution of deletion (I) or insertion (J) events per allele and total insertion length per allele (K).

(L) Insertion frequency of all four DNA nucleotides in edited alleles generated by Cas9-TdT.

(Figure 2H). There were also more insertions across the 10 target sites in DARLIN-v0 mice (Figures 2I and S1J), with all four nucleotide identities well-represented within these inserted sequences (Figure S1H). These data demonstrate that compared with Cas9/CARLIN, the DARLIN-v0 mouse line achieves a larger fraction of rare alleles and greater barcode diversity due to more insertions and fewer large-scale deletions enabled by the Cas9-TdT editing system.

DARLIN-v0 mouse line works across tissues and yields >1 million alleles

We next sought to demonstrate that the DARLIN-v0 mouse can work across tissues. We examined the editing patterns of the target array in intestine, kidney, lung, liver, spleen, and gonad via bulk RNA-seq (Figure 2B). The editing efficiency was >90% across these tissues as compared with 30%–50% as reported for Cas9/CARLIN,[19] with only ∼4% background editing without Dox induction (Figure 2J). The singleton-allele fraction was ∼70% and was comparable across tissues over a broad range of observed allele numbers (Figure 2K). Pooling alleles from all tissue samples gave a similar singleton-allele fraction, suggesting that individual tissue samples were dominated by distinct alleles. Indeed, within the same mouse, only 5% of lineage barcodes were shared between the spleen, kidney, and intestine (∼30,000 alleles each), indicating that most alleles were relatively rare (Figure 2L).

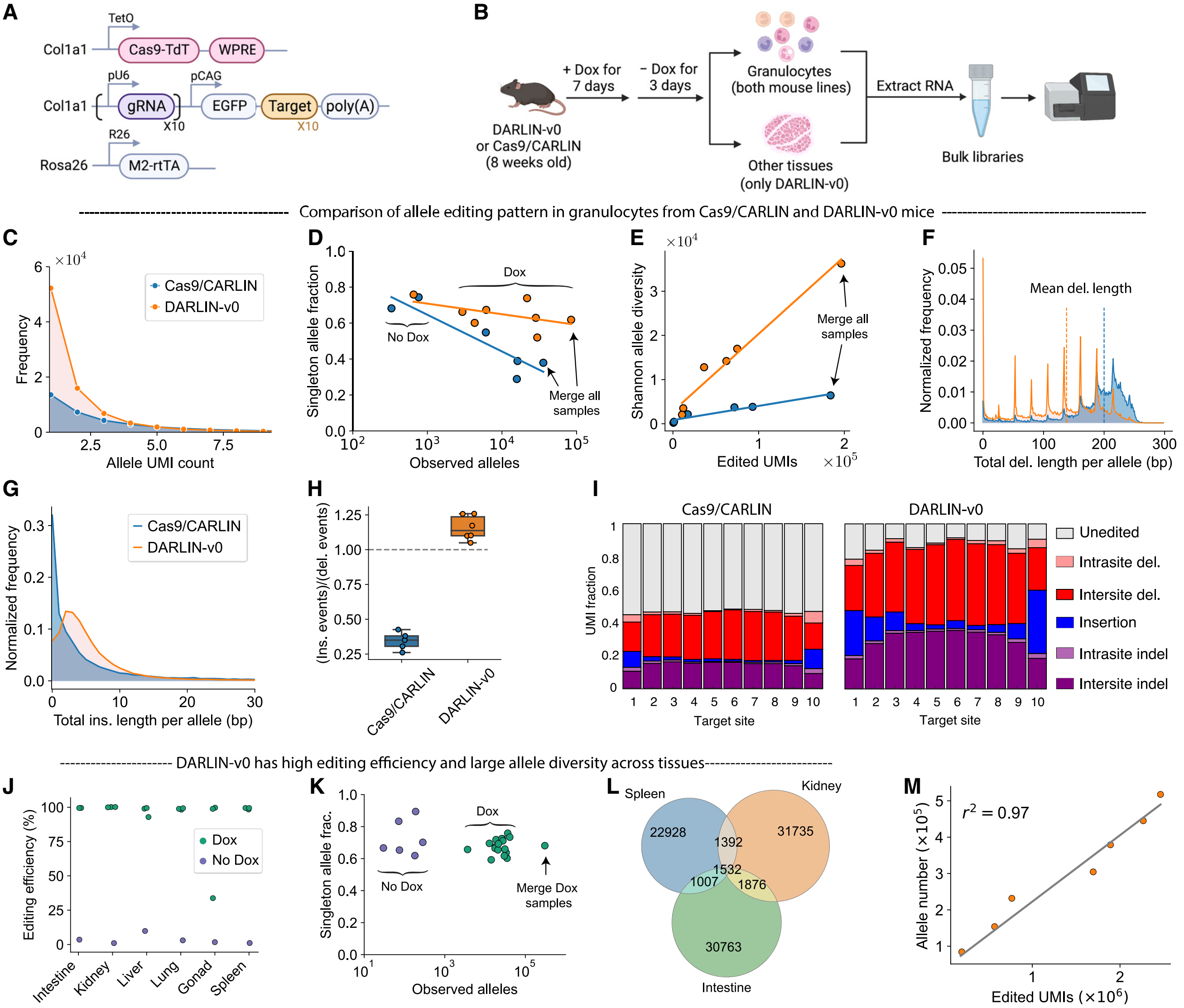


Figure 2. Characterization of DARLIN-v0 mice

(A) Schematic of DARLIN-v0 system. The mutant tetracycline reverse transactivator (M2-rtTA) expressed from the *Rosa26* locus activates Cas9-TdT-WPRE expression upon Dox administration, leading to editing in the target array. WPRE: woodchuck hepatitis virus posttranscriptional regulatory element.

(B) Experimental scheme to compare the target-array editing between DARLIN-v0 and Cas9/CARLIN.

(C–I) Comparison of the alleles generated in granulocytes of Cas9/CARLIN with those of DARLIN-v0.

(C) Histogram of UMIs per allele.

(D) Singleton-allele fraction as a function of observed alleles. Each point represents a mouse replicate except for the rightmost points.

(E) Shannon allele diversity as a function of total UMI counts (edited cells).

(F and G) Distribution of total deletion length (F) or total insertion length (G) per edited allele.

(H) Box plot of the insertion-to-deletion ratio.

(I) Relative UMI fraction of editing patterns across ten target sites.

(J–L) Evaluation of target-array editing in multiple tissues from DARLIN-v0.

(J) Observed editing efficiency.

(K) Singleton-allele fraction.

(L) Venn diagram of the allele overlap between three tissues.

(M) Observed alleles as a function of edited UMIs in DARLIN-v0. These points correspond to merged samples from increasing mouse tissues and their replicates.

We next estimated the clonal barcode diversity of the DARLIN-v0 model. By progressively pooling alleles from an increasing number of mouse tissues and their replicates, we observed that the number of distinct alleles increased linearly with the number of edited cells (goodness of linear fit $r^2 = 0.97$; Figure 2M), suggesting little-to-no saturation. Pooling alleles across all available data from the DARLIN-v0 mouse line, we observed 5.2×10^5 unique alleles in total (Figure 2M), with a singleton fraction of 0.62 (Figure S1K). The large fraction of singletons implied that many more unobserved alleles would be detected if we sample more cells.[40] Because our sampling was far from saturation, we could not directly calculate the maximum number of alleles that could be produced by the DARLIN-v0 mouse. We therefore inferred the number of total alleles using the Chao1 estimator.[40,41] This approach yielded an estimate of 1.3×10^6 possible alleles, a value at least 30 times greater than the reported 44,000 total alleles reported for Cas9/CARLIN.[19] To conclude, our DARLIN-v0 mouse line is suitable for lineage tracing in various tissues and has a large barcode capacity.

DARLIN mice contain three independent target arrays

To further increase the clonal barcode diversity for organism-wide lineage tracing, we generated two additional mouse lines, each with a distinct target array: one was integrated at the *Tigre* locus (yielding the *Tigre* target-array [TA]) and the other at the *Rosa26* locus (*Rosa26* target-array [RA]). Both TA and RA reused the same 10 target sequences from the original CA, such that they would be edited by the existing 10 tandem gRNAs, but in different orders (Figure S2A). We generated an additional tetO-Cas9-TdT knockin mouse line carrying an independent copy of the 10 gRNAs used in the original CARLIN construct, with the goal of enhancing barcode editing. By crossing homozygous mice having all three target arrays with homozygous *Col1a1*$^{\text{tetO-Cas9-TdT-gRNA/tetO-Cas9-TdT-gRNA}}$: *Rosa26*$^{\text{M2-rtTA/M2-rtTA}}$ mice, we obtained DARLIN-v1 mice (Figure 3A). Unless otherwise stated, all the data presented below were generated from this DARLIN-v1 line, referred to hereafter simply as DARLIN mice.

To evaluate the editing performance across the CA, TA, and RA loci, we induced six adult DARLIN mice with Dox for 1 week and analyzed the alleles from bone-marrow granulocytes with bulk DNA sequencing (Figure 3B). We found that the three target arrays achieved a similar editing efficiency (Figure 3C), comparable Shannon allele diversity (Figures 3D and S2B), as well as ~60% singleton-allele fraction (Figure 3E). We further confirmed that the different arrays had similar editing patterns (Figures 3F and S2C–S2E). These data were in agreement with the above CA data from the DARLIN-v0 mouse (Figures 2C–2I). We also observed that CA, TA, and RA had similar editing dynamics upon Dox induction in embryos, and they were similarly saturated at ~100% after 24 h of Dox treatment (Figures S2F and S2G). We conclude that in DARLIN, TA and RA perform comparably to the original *Col1a1*-based target array with respect to editing efficiency and barcode diversity.

We also confirmed that the editing of the three arrays was independent of each other (Figure S2I). This implies that the maximum theoretical number of barcodes in DARLIN could reach $\sim 10^{18}$, assuming each locus can generate at least 10^6 alleles. Notably, this barcode complexity far exceeds the total number of cells ($\sim 10^{10}$) in an adult mouse.

To identify rare alleles that can uniquely label a clone, we performed an experiment to measure intrinsic allele frequencies (Figure S3). We collected an allele bank with $\sim 10^5$ alleles for each of the three arrays, aggregated across three bulk DARLIN mouse replicates. We inferred the generation probability ρ for each of these alleles, which we then used to identify alleles that were statistically unlikely to be contaminated by barcode homoplasy and therefore likely to have labeled real clones in other experiments within this study. At a false discovery rate (FDR) of 0.01 for reliable clone identification, we estimated that the alleles from our bank can label $\sim 10^4$ reliable clones when considering only one target array and $\sim 10^{12}$ clones if all three arrays were to be used in combination (Figure S3D; STAR Methods). In practice, the barcoding capacity can be greatly expanded by including *de novo* alleles (i.e., alleles not found in our relatively small allele bank). In fact, ~80% of alleles observed in our datasets from actual lineage-tracing experiments were *de novo* alleles (Figure S5D).

DARLIN achieves superior single-cell lineage coverage

Because target arrays in DARLIN are transcribed, one can simultaneously profile the lineage barcodes and transcriptomes of single cells. Transcriptomic information enables systematic resolution of cell states, which is crucial for understanding lineage relationships in a heterogeneous population without conventional sorting markers. However, only cells with lineage barcodes that are detected, edited, and rare will be useful for downstream lineage analysis to avoid barcode homoplasy (Figure 3G). We therefore systematically evaluated the characteristics of single-cell lineage tracing in the DARLIN mouse line.

We labeled DARLIN mice at E17.0 and generated a single-cell lineage-tracing dataset of blood cell progenitors by sorting $\text{Lin}^-\text{cKit}^+$ cells from skull bone marrow (Figure 3H). We obtained 6,094 cells after quality control (QC) (Figure 3I) and detected alleles from at least one target locus in 81% of cells, half of which (overall 40%) contained at least two of three target loci (Figure 3J). The target array at the *Tigre* locus exhibited more efficient capture due to its higher expression (Figures 3J and S2H). Among these 6,094 cells, 3,839 cells (63%) had at least one rare allele (Figure 3K), and this fraction was comparable across the seven blood cell types in the skull (Figure 3L). We profiled hematopoietic cells from four other tissues (left-leg-derived bone marrow, liver, lung, and spleen) in single-cell assays (Figure 3H) and also observed ~60% of cells with at least one rare allele (Figure 3M).

Next, we systematically compared the single-cell lineage readout between DARLIN and Cas9/CARLIN[19] mouse lines. On average, we detected expression from at least one lineage locus in 80% of cells derived from the DARLIN mouse, compared with ~50% from Cas9/CARLIN (Figure 3N). Among the detected lineage barcodes, ~80% were further edited in DARLIN, compared with ~35% in Cas9/CARLIN, which was consistent with our observations of editing efficiency in mouse embryos (Figure S2J) and adult mice (Figure 3C). Finally, among the edited cells, ~93% of cells from DARLIN mice had at least one rare allele, whereas this fraction was only 55% for Cas9/CARLIN mice. In total, DARLIN mice achieved ~60% of cells passing all three filters (i.e., had an allele that was detected, edited, and rare), compared with ~10% in Cas9/CARLIN. Our assessment of Cas9/CARLIN agreed with our previous assessment[19] and was consistent across two additional single-cell Cas9/CARLIN datasets analyzed with our method (Figure 3N). Together, the above data demonstrate that the DARLIN mouse line has superior

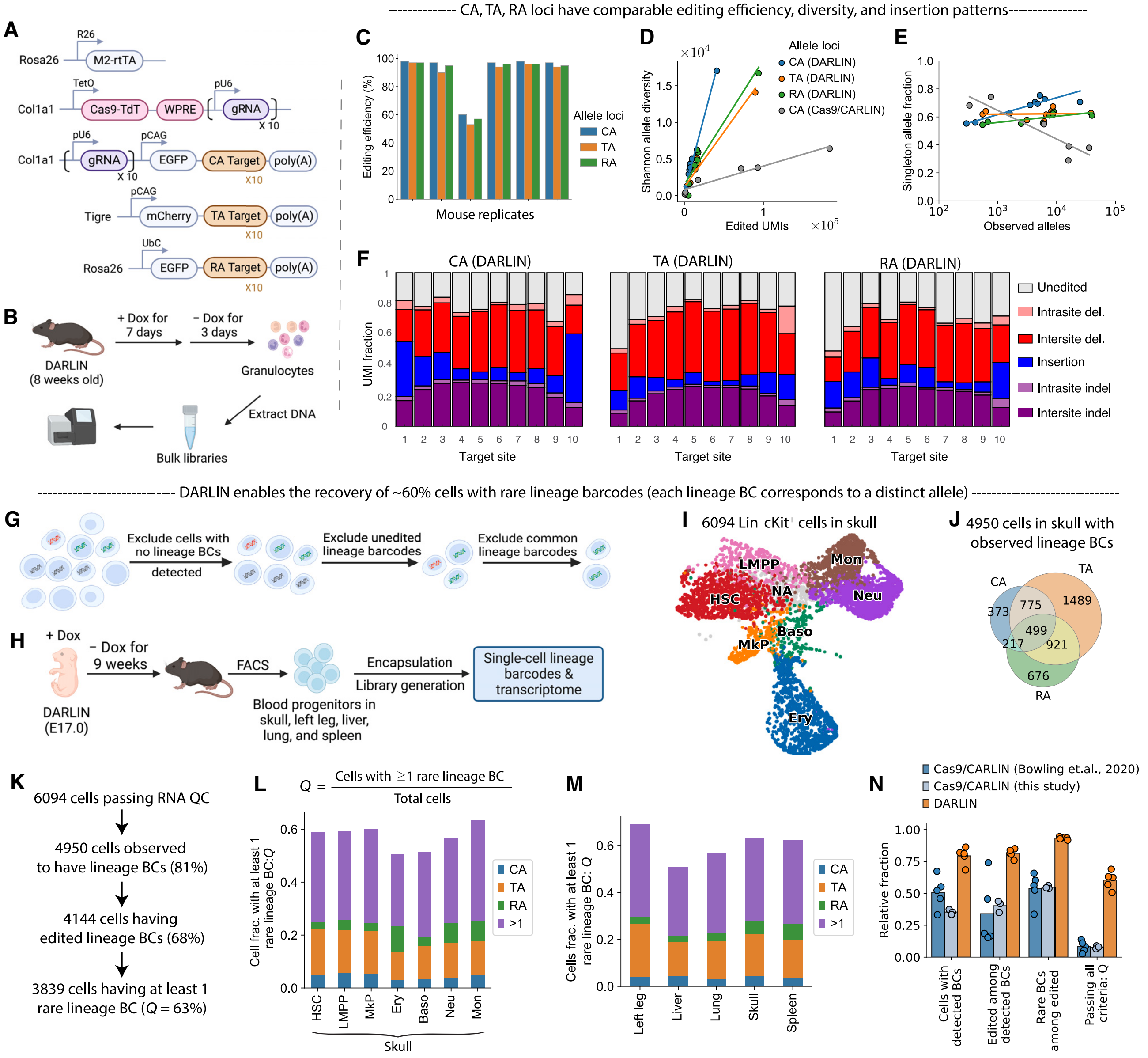


Figure 3. Characterization of DARLIN mice

(A) Genetic elements of the DARLIN system.

(B) Experimental scheme to compare the target-array editing in the three loci of DARLIN.

(C–F) Analyses of alleles generated from the three loci: editing efficiency (C), Shannon allele diversity as a function of detected UMIs (D), the singleton-allele fraction as a function of observed alleles (E), and frequencies of editing patterns across the 10 target sites (F).

(G) Quality control (QC) pipeline for selecting single cells with reliable lineage barcodes.

(H) Single-cell lineage-tracing experimental design with DARLIN.

(I) UMAP embedding of the transcriptomes for the skull-derived Lin$^-$cKit$^+$ cells. HSC, hematopoietic stem cell; LMPP, lymphoid-biased multipotent progenitor; MkP, megakaryocyte progenitor; Ery, erythrocyte; Baso, basophil; Neu, neutrophil; Mon, monocyte. UMAP, uniform manifold approximation and projection for dimension reduction.

(J) Venn diagram showing the number of cells for which each type of target array or combination of these was detected.

(K) Cell number at each filtering step for the skull dataset.

(L and M) Fraction of cells for which a rare allele was detected from at least one target array, either for different cell types from the skull (L) or blood cells from different tissues (M). Each bar is colored according to the percentage of cells with alleles at only a single locus or multiple loci (>1).

(N) Fraction of cells that passed each QC step (described in G) between DARLIN and Cas9/CARLIN. The DARLIN data are from (M), the Cas9/CARLIN data generated in this study were collected from the head, tail, and trunk of a mouse embryo, and the published Cas9/CARLIN data (collected from bone marrow) correspond to those of Figure 6 from Bowling et al.[19]

single-cell lineage coverage and a barcode diversity that exceeds the number of cells in an entire adult animal.

Mapping cell-fate choices among unperturbed blood progenitors *in vivo*

We next demonstrated the utility of DARLIN to study cell-fate choice during developmental hematopoiesis. Several studies have shown that hematopoietic stem and progenitor cells (HSPCs) can be divided into subpopulations with functional heterogeneity,[10,21,22,42] including subsets with distinct fate biases. However, it is unclear when this fate bias is established during development and what are the molecular features of these biased HSPCs.

We re-analyzed our single-cell lineage-tracing data from skull-derived bone marrow induced at E17.0 and collected in adulthood (Figure 3H). We identified six major cell types among the 6,094 profiled single cells: hematopoietic stem cells (HSCs), lymphoid-biased multipotent progenitors (LMPPs), megakaryocyte progenitors (MkPs), erythrocytes, neutrophils, and monocytes (Figures 3I, 4A, and S4A). We integrated information from the three target-array loci to assign a clone ID to each cell (STAR Methods). In total, we identified 1,034 distinct clones (Figure 4B): some clones occupied multiple cell fates (Figure 4C, left panel), whereas others had only one observed fate outcome (Figure 4C, middle and right panel). With these data, we asked if some HSPCs (HSCs and LMPPs) demonstrated differentiation bias toward specific fates (Figure 4A).

The clonal coupling scores across major cell types (i.e., a normalized correlation to measure how often two cell types jointly appear within the same clone; STAR Methods) suggested a strong lineage coupling between MkPs and HSCs ($p < 0.001$) and between monocytes and LMPPs ($p < 0.05$) (Figures 4D and S4B). This agrees with earlier reports that a subset of HSCs can directly generate MkPs[8,22,43–46] and that LMPPs are primed to generate monocytes rather than neutrophils.[21] In murine hematopoiesis, definitive blood progenitors arise at E10.5 with the formation of *Runx1*-expressing clusters within the aorta-gonad-mesonephros (AGM) region in the embryo.[47] At around E11.5, these progenitors begin to migrate to the fetal liver where they first undergo rapid expansion before colonizing the bone marrow at around the time of birth (i.e., E19–E21). Considering that barcoding was induced at E17.0, a developmental time point when HSCs still reside in the fetal liver, our data suggest that HSCs at this time already carry functional features that will be evident even after their migration to the bone marrow. Thus, MkP bias is likely to arise earlier than what has been previously reported.[48] We found that 48% of our 187 clones that both contained multiple cells and included at least one HSPC had a single clonal fate (Figure 4D), suggesting the possibility of early fate bias. Inspecting those HSPCs that were clonally associated with a single mature fate, we found that only MkP-biased clones had distinct transcriptomic signatures (Figure 4E). We previously developed CoSpar, a computational approach that utilizes coherent and sparse lineage dynamics to robustly infer early cell-fate choice.[49] We applied CoSpar to infer early fate priming by integrating transcriptomic and lineage information. Consistent with the above observations, CoSpar predicted that MkPs originate specifically from a subset of HSCs (Figure 4F, left panel). Interestingly, CoSpar also predicted that monocytes originate predominantly from LMPPs (Figure 4F, right panel), which agreed with our clonal coupling analysis (Figure 4C). Importantly, we failed to infer such early fate bias when down-sampling our DARLIN data to match the frequency of cells with detected, edited, and rare alleles in Cas9/CARLIN data (Figure S4D).

Next, to identify the early transcriptomic signature of MkP-biased HSCs, we inferred the differentiation trajectory from HSCs to MkPs using the above CoSpar predictions, then split the MkP-biased HSCs into two populations based on their pseudotime: early or late MkP bias (Figure 4G). Compared with HSCs without MkP bias, the early MkP-biased HSCs exhibited enriched expression of genes involved in maintaining long-term HSC identity (*Mecom*, *Mllt3*, and *Hlf*), cell-cycle inhibition (*Ifitm1*, *Txnip*, and *Ifitm3*), and megakaryopoiesis regulation (*Tbxas1*, *Mpl*, and *Meis1*) (Figures 4H and 4I).[22] We also identified many genes without an established association with MkP bias in HSCs (Figure 4J), including the transcription factors *Klf12*, *Sox5*, *Rora*, *Pbx3*, *Pbx1*, and *Gata2* (Figure S4C). Taken together, our analyses demonstrated that the DARLIN mouse line generates high-quality single-cell lineage-tracing data that resolves early fate bias within HSCs, leading to the identification of gene signatures of MkP-biased HSCs in unperturbed hematopoiesis *in vivo*.

Lineage relationships of blood cells across bones reveal HSC migration dynamics over development and adulthood

Next, we used the DARLIN mouse line to systematically evaluate clonal dynamics of the migration of hematopoietic progenitors over development and adulthood (Figure 5A). Although it is appreciated that HSCs migrate from the fetal liver to the bone marrow at around the time of birth, the clonality of bone-marrow colonization and the extent of HSPC circulation during ontogeny remain unclear. Similarly, the extent of migration and differentiation in the adult bone marrow remains poorly explored. HSC circulation in adulthood was previously studied in mice by parabiosis.[50–52] In these studies, Wright et al. observed that up to 8% of HSCs migrated from one mouse to the other over 39 weeks[50]; however, a later study observed only 1%–2.5% migratory HSCs.[51] Parabiosis experiments are highly invasive and lead to injury and inflammation, which might influence the behavior of HSPCs in such studies. The high barcoding capacity of the DARLIN model presented us with the unique opportunity to address these questions at the level of individual clones in a completely physiological context.

We induced DARLIN mice at different developmental stages (adulthood, neonate, and E17.0). After 4 months, we dissected bone marrow from four locations (skull, spinal cord, left leg [i.e., femur, tibia, and fibula], and right leg) and used fluorescence-activated cell sorting (FACS) to sort long-term (LT) HSCs, MPPs, myeloid progenitors (MyPs), and MkPs from each bone to profile their lineage barcodes via bulk RNA-seq (Figure 5B, upper panel and S5A). In a separate experiment, we also induced one DARLIN mouse at E10.0, waited 2 months, and profiled the lineage barcodes across major blood cell types sorted from four different bones (Figure 5B, lower panel). In these bulk RNA-seq experiments, a clone is a set of UMIs sharing the same (rare) lineage barcode, which may come from different bones or cell types.

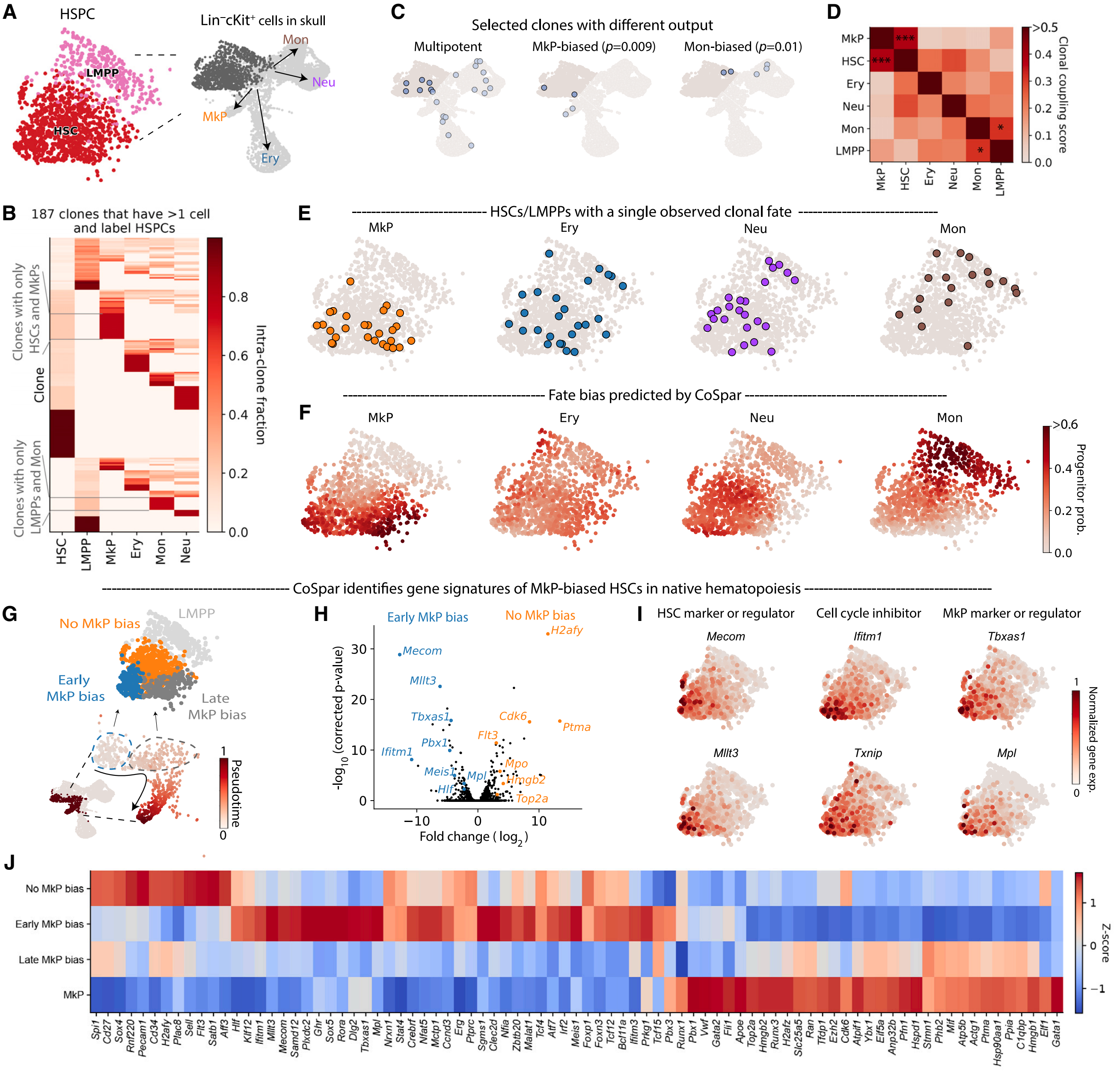


Figure 4. Early fate priming among hematopoietic stem and progenitor cells (HSPCs)

(A) UMAP embedding of HSPCs (see also Figure 3I).

(B) Clonal profile of the normalized proportion of each annotated cell type (column) within each clone (rows). Only the 187 clones labeling HSPCs are shown.

(C) Selected clones with different fate outcomes. p values of clonal fate bias are shown for the latter two clones.

(D) Heatmap of clonal coupling scores across major cell types (STAR Methods). Coupling scores that are statistically significant are indicated (*$p < 0.05$; ***$p < 0.001$).

(E) UMAP embedding of HSPCs, highlighting cells that were clonally associated with a single mature cell type.

(F) CoSpar-predicted probability of each HSPC to generate a mature cell type.

(G) Identification of early MkP-biased HSCs with CoSpar.

(H) Volcano plot of differentially expressed genes when comparing early MkP-biased HSCs with inferred HSCs with no MkP bias.

(I) UMAP embedding of HSPCs overlaid with expression of selected genes.

(J) Heatmap showing the expression of selected genes across different HSPC clusters and MkP. *Z* scores were calculated per gene within the four cell populations.

Please cite this article in press as: Li et al., A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells, Cell (2023), https://doi.org/10.1016/j.cell.2023.09.019

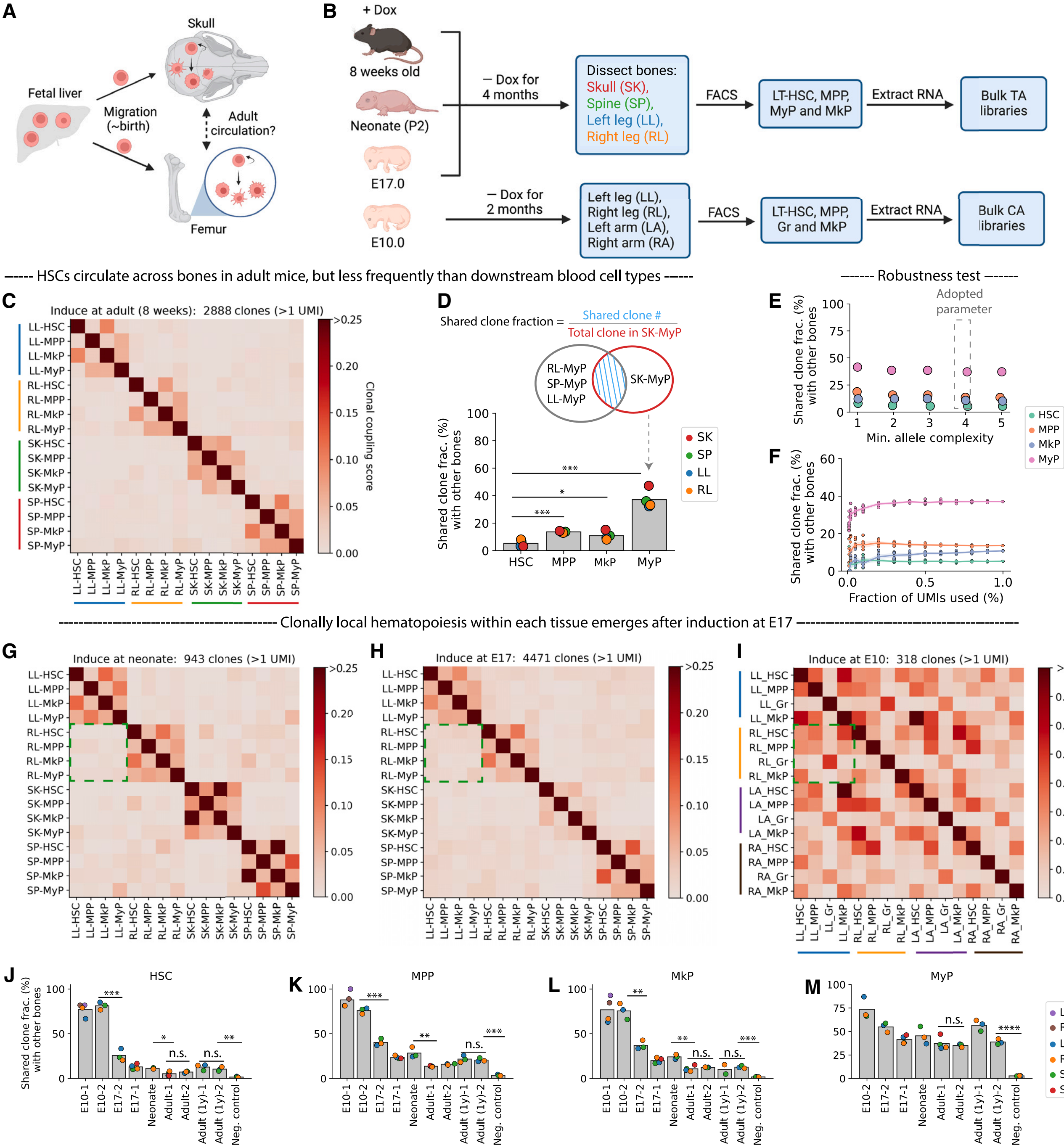


Figure 5. Lineage relationships of blood cells across bones

(A) Schematic of migration of blood progenitors across developmental stages.

(B) Experimental design to investigate HSC migration dynamics. Leg bones include the femur, tibia, and fibula, whereas arm bones include the humerus, ulna, and radius.

(C–F) Clonal analysis of bulk lineage-tracing data from week-8 induction.

(C) Heatmap of clonal coupling scores between cell types from each bone. (D) Shared clone fraction of each cell type across bones (*p < 0.05; **p < 0.01; ***p < 0.001; ****p < 10^{-4}, t test).

(E) Shared clone fraction for each cell type over different thresholds of minimum allele complexity for excluding low-complexity alleles.

(F) Shared clone fraction when performing different amounts of UMI down-sampling.

(legend continued on next page)

First, we determined to what extent hematopoietic progenitors circulate across different bones during adulthood by analyzing mice induced at 2 months of age. The presence of a clone in more than one bone indicates inter-bone migration, by which an individual HSC (or progenitor) divides and colonizes a different bone-marrow niche. Calculating the clonal coupling scores between all pairs of cell types from all sorted populations, which accounts for clone identities and their sizes, we observed that hematopoietic populations were strongly related in clonal origin within each bone but not between bones (Figure 5C). This is consistent with the idea that hematopoiesis is predominantly maintained locally within each bone in the adult, at least within 4 months. Indeed, each of the clones resided predominantly in one bone, with only a small fraction of cells (UMIs) detected in other bones (Figure S5B). Considering the fraction of HSC-containing clones found in one bone (e.g., skull HSCs) that are also detected in HSCs from other bones (irrespective of clone size), we observed that ~5% of HSC-containing clones were shared with HSCs from at least one other bone (Figure 5D). The overlap fraction increased significantly to ~14% for MPPs (t test, $p < 10^{-3}$) and to ~40% for MyPs ($p < 10^{-3}$) (Figure 5D). To exclude contamination from common alleles, we only used *de novo* alleles (~80% of all alleles) from this experiment that were not found in our pre-assembled allele bank with ~100,000 alleles (Figure S5D). Accordingly, the inferred shared clone fraction was robust to (1) the mutational complexity of the alleles considered (Figure 5E), (2) down-sampling of UMIs (Figure 5F), and (3) read cutoffs used for allele calling (Figure S5E). We also evaluated the extent of HSPC migration with age. Extending the chase period from 4 months to 1 year increased the observed shared clone fraction in HSCs from ~5% to ~12% (Figure 5J), suggesting that HSPC migration occurs at a low level and accumulates with age. Overall, our data extend the earlier findings of HSC circulation between bones[50,51] and provide definitive evidence that this process actively occurs in a native physiological context. Our findings also demonstrate that less primitive populations like MPPs and MyPs circulate more actively.

We next studied the dynamics of inter-bone migration in the neonate. It is unclear whether the migration of post-birth HSCs is more dynamic than those in the adult. Our results demonstrated that at 4 months after birth, overall local hematopoiesis was still prevalent even when barcoding was induced in the neonatal stage (Figure 5G). There was, however, ~11% shared clone fraction of HSCs between the bones studied (Figure 5J), higher than the ~5% when induced in adulthood (p = 0.01). Thus, our results suggest an increased rate of inter-bone HSC migration post birth in comparison to adulthood, consistent with an indirect study based on live imaging.[53]

We also induced at E17.0, a stage when HSCs are predominantly located in the fetal liver.[54] Labeling at this stage will likely result in effective barcoding right before birth, thereby minimizing the effects of clonal expansion before migration. Remarkably, we still observed predominantly local hematopoiesis across bones 4 months later (Figures 5H and S5B). To confirm that we could detect delocalized hematopoiesis in DARLIN, we labeled embryos at E10.0, with the goal of labeling clones that would further expand in the fetal liver and colonize multiple bones. In this context, as expected, the clonal coupling scores of blood cells within the same bone were comparable to those between different bones (Figure 5I), and an ~80% shared clone fraction between bones was observed for different cell types (Figures 5J–5M). Thus, our observations suggest that HSCs labeled in the late fetal liver stages (E17.0) predominantly seed one single bone microenvironment and proliferate and differentiate locally after birth.[54] The shared clone fractions from induction at E17.0 were similar to those of induction at neonate but still higher than those of induction in adulthood for HSCs, MPPs, and MkPs (Figures 5J–5L). We also observed that MyPs had comparable shared clone fractions when labeling across these stages (Figure 5M). This difference was consistent with MyPs undergoing more active circulation in the adult stage.

Importantly, when we induced barcoding in adult mice for only 3 days and then immediately profiled alleles, we observed only a ~1% shared clone fraction across cell types (Figures 5J and S5C). This suggests that technical issues (i.e., background barcoding) have minimal effects on our observations. Additionally, our results were corroborated by performing independent analyses using barcodes amplified from the CA and RA loci within the same mice (Figure S5F). In conclusion, the high barcoding capacity of the DARLIN model has allowed us to obtain unique insight into the process of HSC migration during development and adulthood.

Camellia-seq simultaneously profiles chromatin accessibility, DNA methylation, gene expression, and lineage information in single cells

Integrating lineage tracing with single-cell transcriptomic measurement enables systematic dissection of fate biases for a transcriptomically heterogeneous population.[20–22,49] The epigenetic state of a cell also plays a crucial role in regulating its dynamics and function.[23–25] An integrative measurement of lineage, transcriptome, and epigenome at the single-cell level would enable a deeper understanding of how cell-fate choice is regulated and how cell identity is maintained across different modalities. Here, we developed a sequencing method to simultaneously measure chromatin accessibility, DNA methylation, gene expression, and lineage information in single cells (Camellia-seq) (Figure 6A). Camellia-seq extends scNMT-seq[55–58] by incorporating lineage barcode measurement. Briefly, a single cell is split into nuclear and cytoplasmic fractions. Endogenous mRNAs and expressed lineage barcode transcripts are reverse-transcribed and amplified from the cytoplasmic fraction via a modified STRT-seq protocol.[59,60] The nuclear fraction is treated with GpC methyltransferase, which preferentially methylates cytosine from GpC dinucleotides within regions of open chromatin.[61] The

(G–I) Heatmap of clonal coupling scores between cell types across bones for mice induced at the neonate stage (G), E17.0 (H), and E10.0 (I).
(J–M) Shared clone fraction with other bones when editing was induced at different developmental stages for HSC (J), MPP (K), MkP (L), and MyP (M). When present, "−1" and "−2" indicate data from replicate mice. For the E10.0, E17.0, neonate, and adult samples, the −Dox waiting time durations were as described in (B), and for the adult (1 year) samples, they were 1 year, and the negative control samples were induced in adulthood and immediately profiled.

Please cite this article in press as: Li et al., A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells, Cell (2023), https://doi.org/10.1016/j.cell.2023.09.019

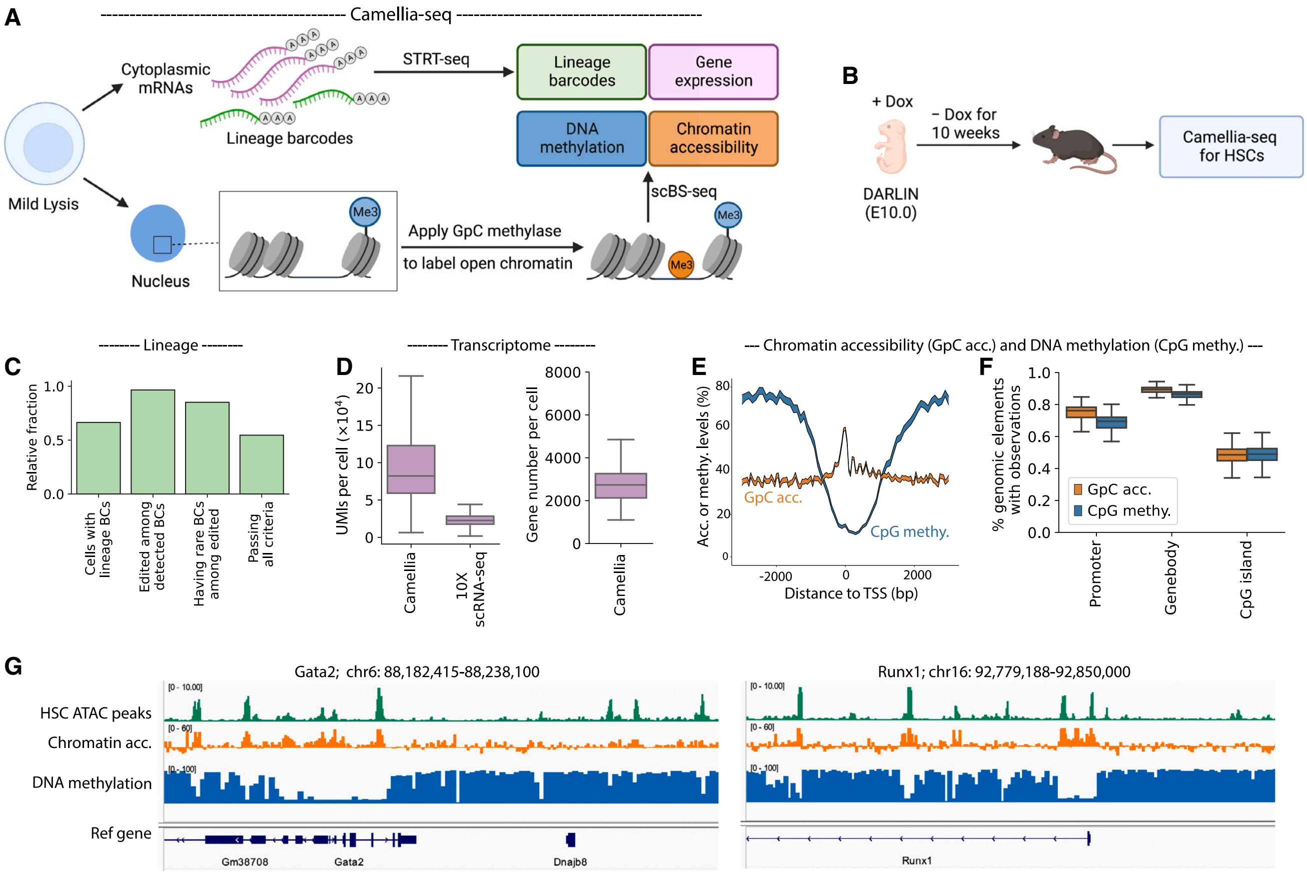


Figure 6. Joint profiling of lineage, gene expression, chromatin accessibility, and DNA methylation with Camellia-seq

(A) Schematic of Camellia-seq.

(B) Experimental scheme to profile bone-marrow HSCs with Camellia-seq.

(C) Fraction of cells that passed each QC step described in Figure 3G.

(D) Box plots showing the number of observed UMIs (left) or genes (right) per cell for the scRNA-seq data generated with Camellia-seq. The corresponding UMIs count per cell from the 10× Genomics protocol (Figure 3I) is also shown.

(E) The average chromatin-accessibility or DNA-methylation profile over the transcription start sites (TSSs) of ∼20,000 different genes in a cell.

(F) Box plot showing the genomic coverage across promoters, gene bodies, and CpG islands. Each GpC site must be covered by ≥3 reads, and CpG site by ≥ 1 read.

(G) Pseudobulk chromatin accessibility and DNA methylation surrounding the TSS of *Gata2* and *Runx1*. The bulk HSC ATAC-seq peaks from Li et al.[54] are also shown. ATAC-seq, assay for transposase-accessible chromatin with sequencing.

endogenous DNA methylation (methylated cytosine in CpG dinucleotides) and accessible chromatin (methylated cytosine in GpC dinucleotides) are then profiled with single-cell bisulfite sequencing.[62]

We profiled HSCs with Camellia-seq to evaluate the quality of each data modality. We induced lineage labeling in the DARLIN mouse at E10.0, when HSCs have just formed in the AGM region, and extracted HSCs (Lin$^-$cKit$^+$Sca1$^+$CD48$^-$) from 9-month-old adult bone marrow to perform Camellia-seq (Figure 6B). Approximately 50% of the single cells profiled with Camellia-seq had a rare lineage barcode (Figure 6C). We observed a median transcriptomic abundance of ∼100,000 UMIs derived from ∼3,000 genes per cell (Figure 6D). Using epigenomic modalities from single cells, we reproduced the stereotypic pattern that DNA methylation decreases within 1 kb of the transcription starting site (TSS), whereas the chromatin accessibility is greatest near the TSS and decreases with an oscillatory pattern in the direction of transcription initiation[56] (Figure 6E). Furthermore, Camellia-seq achieved a high genomic coverage: ∼70% of promoters and ∼90% of the gene bodies were represented with at least 3 detected GpC sites and 1 CpG site (Figure 6F). By aggregating single-cell epigenomic measurements into a pseudobulk dataset, we further confirmed that the resulting chromatin-accessibility measurements largely agreed with bulk ATAC-seq measurements of HSCs from a published dataset[54] (Figure 6G; Pearson $r = 0.63$ for displayed regions) and anti-correlated with DNA-methylation measurements around promoters in our data (Figures 6E and 6G).

DNA methylation maintains strong clonal memory of HSCs over time

We utilized Camellia-seq to gain insight into the question of molecular memory in cell lineages: do cells from the same clonal

Please cite this article in press as: Li et al., A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells, Cell (2023), https://doi.org/10.1016/j.cell.2023.09.019

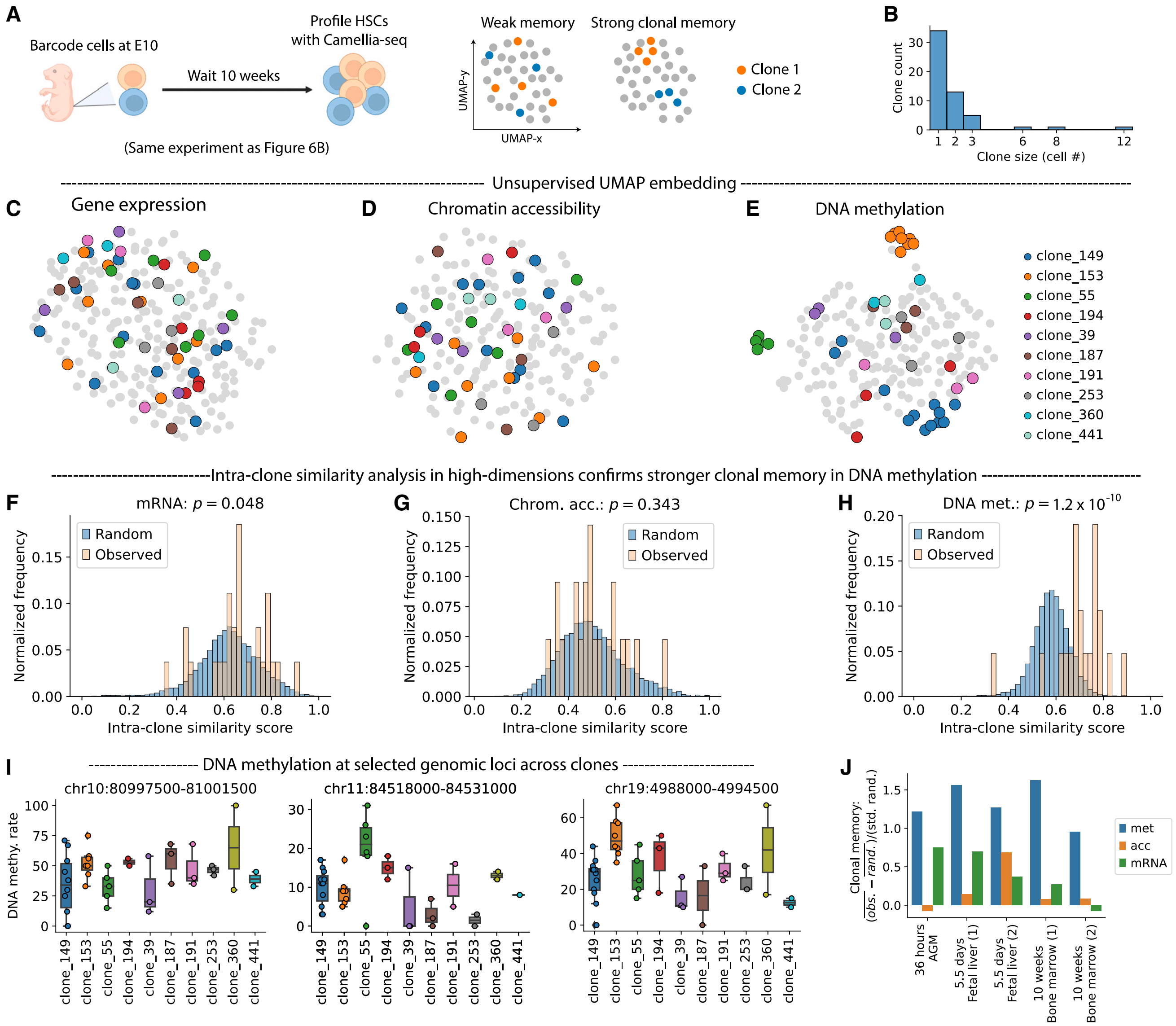


Figure 7. Transcriptomic and epigenomic memory of HSCs within each clone

(A) Experimental design (left) and classification of clonal memory (right).

(B) Distribution of clone sizes from profiled HSCs.

(C–E) UMAP embedding generated using either gene-expression (C), chromatin-accessibility (D), or DNA-methylation (E) data from Camellia-seq. Cells are colored by their clone identities.

(F–H) Distribution of intra-clone similarity scores from 21 observed and 21 × 1,000 randomized clones, calculated using either gene expression (F), chromatin accessibility (G), or DNA methylation (H). p values for each modality were calculated using the Wilcoxon rank-sum test.

(I) DNA-methylation levels at selected genomic loci across the top 10 largest clones.

(J) Clonal memory score for each modality across all mouse samples collected at different developmental stages.

lineage retain molecular signatures indicating that they arose from the same founder cell, despite potential changes in the external cellular environment among the daughter cells from the same clone? To exclude confounding factors like cell differentiation, we focused on purified HSCs and profiled them from the adult bone marrow with Camellia-seq, following Dox induction at E10.0 (Figure 7A, left panel). One hypothesis is that HSCs within the same clone are indistinguishable from those of other clones when assessing their genome-wide molecular states in an unbiased manner, a scenario for weak clonal memory, and the opposite case is that HSCs are more similar within the same clone than across clones, a scenario of strong clonal memory (Figure 7A, right panel). We sought to address this problem with our data and measured cell-cell similarity within individual clones with respect to gene expression, chromatin accessibility, and DNA methylation.

Please cite this article in press as: Li et al., A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells, Cell (2023), https://doi.org/10.1016/j.cell.2023.09.019

We restricted our analysis to clones containing ≥2 cells (21 clones in total) (Figure 7B). For each of the three molecular modalities, we separately performed unsupervised dimensionality reduction and visualized the results via UMAP[63] embedding (STAR Methods), overlaying the clone identities on the embedding. For the 10 largest clones, individual cells within the same clone were largely scattered across the embeddings generated using either gene expression or chromatin accessibility (Figures 7C and 7D), suggesting weak clonal memory with respect to these two modalities. Conversely, cells belonging to the same clone were strongly co-localized within the embedding generated from DNA methylation, suggesting stronger clonal memory (Figure 7E). To quantify the strength of clonal memory with respect to each modality, we calculated the similarity between any two cells using the Pearson correlation coefficient and compared the average similarity within the same clone with those from randomized clones having the same clone size distribution. These analyses demonstrated that clonal memory with respect to gene expression was barely significant ($p = 0.049$; Wilcoxon rank-sum test; Figure 7F), and chromatin accessibility was not significant at all ($p = 0.34$, Figure 7G). However, the clonal memory with respect to DNA methylation was highly significant ($p = 1.2 \times 10^{-10}$; Figure 7H). Because most of the observed clones were small (i.e., 2–3 cells), we also evaluated each clone individually, finding that 19 out of the 21 tested clones had significant intra-clone similarity with respect to their DNA methylation (Figure S6A). We identified 279 genomic regions with differential CpG methylation among clones ($p < 0.05$; Benjamini-Hochberg-adjusted one-way ANOVA; Figure S6B). We provide three examples in Figure 7I, in which each clone has a different extent of DNA methylation in selected genomic regions. These 279 differential methylated regions were neither located near differentially regulated genes (Figure S6C) nor preferentially associated with any gene-ontology terms (Figure S6D), suggesting they represented random genetic loci rather than functionally relevant regions.

We validated our findings of clonal memory with additional biological samples via Camellia-seq. These include HSCs that were labeled with lineage barcodes for 36 h (AGM HSCs), 5.5 days (fetal liver HSCs, two replicates) (Figure S6E), and an additional replicate of HSCs traced for 10 weeks. In our combined five datasets (mice), ∼750 cells successfully passed QC, comprising a total of 63 clones with ≥2 cells. We confirmed that chromatin-accessibility, DNA-methylation, and gene expression measurements captured stage-specific biological signals (Figures S6F–S6H),[54] and the editing efficiencies were close to 100% in each mouse sample, with ∼55% of cells having valid lineage coverage (Figures S6I and S6J). In these samples, we corroborated our findings that the memory scores were significant with respect to DNA methylation for each of the HSC stages (Figure S6K) and were consistently higher than those from the other two modalities (Figure 7J). Thus, we conclude that DNA methylation can retain the memory of clonally related cells much better than either gene expression or chromatin accessibility.

DISCUSSION

Here, we describe DARLIN, a lineage-tracing mouse line with a superior lineage barcoding capacity and enhanced single-cell lineage coverage. Building on the current Cas9/CARLIN lineage-tracing system, we first incorporated TdT to increase insertion events in lineage barcodes and subsequently expanded DARLIN to include three independent lineage-recording loci. DARLIN can theoretically generate an estimated 10^{18} unique lineage barcodes, has ∼90% barcode editing in the embryo, and allows for ∼80% barcode capture in traditional single-cell assays, leading to ∼60% of profiled cells having rare barcodes for downstream clonal analysis. This translates into more useful clones per sample, more cells per clone, and dramatically reduced experimental costs for generating a dataset with sufficient clonal information to address a biological question. Finally, lineage barcoding in DARLIN can be induced at any time point and across a wide range of tissues, and DARLIN is a stable and genetically defined mouse line that can be shared across a wide biological community.

The massive barcode diversity generated by DARLIN not only increases the fraction of rare clones in our data but also enables the study of large biological systems, such as adult tissue homeostasis, inflammation response, and tissue injury and repair. In many applications, profiling alleles from a single target-array locus may already provide sufficient lineage information, with measurements from the remaining loci providing additional robustness.

The superior performance in DARLIN mice is likely due to the improved genetic design (Figure 3A). The higher allele diversity, apart from having three target arrays, is mainly due to the increase of insertions from TdT and may also benefit from fewer large-scale deletions that result in more degenerate alleles (Figures 2F, S7A, and S7B; see Figures S7C–7E for our proposed model to explain this observation). The enhanced editing results from both the increased expression of Cas-TdT due to inclusion of the woodchuck hepatitis virus posttranscriptional regulatory element (WPRE) (Figure S7F) and also increased conversion of Cas9-induced single-site cleavage into edited alleles with insertions (rather than restoration of unedited sequences by direct blunt-end joining; Figures S7C–S7E). In practice, we have observed higher variability of editing efficiency when inducing adult mice (Figures S1I and 3C) than embryos (Figures S2J and S6I), suggesting the need to further improve the protocol for barcode induction in adult mice.

A high editing efficiency also helps to mitigate the impact of background editing without Dox treatment. We observed ∼4% background editing in an adult 8-week-old mouse (Figure 2J). These alleles resulting from background editing could be further edited at the time of Dox induction due to the high editing efficiency of DARLIN and split into different sub-alleles that label clones at the correct timing (Figures S7G and S7H). Combined with our rare-allele filtering strategy, we have achieved a ∼1% false inter-bone clonal sharing (background noise) in a negative control experiment of adult HSC circulation (Figures 5J–5M and S5C).

In our applications, we demonstrated first that the DARLIN mouse line enables the study of early fate bias within native HSCs at high resolution, leading to the identification of multiple new genes correlated with MkP bias in HSCs. In our second application, we studied the lineage relationship of hematopoietic cells across different bones. Our data demonstrate HSC inter-bone migration in a completely physiological context, with a ~5% shared clone fraction of HSCs between different bones accumulated over 4 months after induction in adulthood and higher fractions in aged animals. These observations support the idea that HSCs continuously circulate at low levels in adulthood.[50] Considering that we dissected only ~70% of the mouse bone marrow, we likely underestimated the extent of HSC migration. Our data also speak to the prevalence of local hematopoiesis even when induction was done either in the late fetal liver or the neonate. Thus, our findings suggest limited migration even after initial bone settlement. Barcoding models such as DARLIN represent a novel approach to study cellular migration in the bone marrow.

In parallel, we have established Camellia-seq to simultaneously profile lineage barcodes, chromatin accessibility, DNA methylation, and gene expression in single cells. Using DARLIN, we showed that Camellia-seq generates high-quality data for each of the modalities. By focusing on HSCs that can self-renew, we demonstrated that genome-wide DNA-methylation patterns, but not chromatin-accessibility or gene expression patterns, stably propagate within individual clones over multiple cell divisions. Finally, our ~750 HSCs profiled via Camellia-seq cover key developmental stages of hematopoiesis and will be a valuable resource to further understand hematopoiesis and, more generally, the interplay between different modalities at the single-cell level.

Additionally, Camellia-seq is compatible with any lineage-tracing approach where the lineage barcodes are transcribed as mRNA.[31] DARLIN mice may also be induced with Dox over a series of time points to generate alleles with a hierarchical structure to obtain more hierarchical cellular lineage relationships of large numbers of cells during tissue development or homeostasis. The lineage barcodes in DARLIN may also be resolved spatially to understand spatial lineage dynamics in tissues. Overall, the DARLIN mouse line and Camellia-seq method provide a powerful tool for studying the relationships and underlying molecular mechanisms of diverse biological processes.

Limitations of the study

The target arrays in DARLIN still suffer from array deletions, which might limit their use for lineage reconstruction across multiple cell divisions. This may be circumvented by generating sequential mutation events along the recording array to produce hierarchically labeled clones.[64] Camellia-seq is currently a low-throughput and costly plate-based method that requires deep genomic coverage for each cell. A cost-effective and high-throughput method would be desirable.

STAR★METHODS

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

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2023.09.019.

ACKNOWLEDGMENTS

We are grateful to members of F.D.C. laboratory; the IDDRC Gene Manipulation Core (funded by NIH P50 HD105351) for ESC injection and chimera generation; Ronald Mathieu, Mahnaz Paktinat, Ranjan Maskey, and Betelhem Gemechu from HSCI-BCH Flow Cytometry Research Lab for guidance and assistance with FACS; Alejo Rodriguez-Fraticelli for assistance with designing the Cas9-TdT plasmid; Sachin Patel for assistance with designing the clonal memory experiment; Wei-Chien Yuan for assistance with hydrodynamic injections and IP injections in mice; Fan Zhou and Chloé Baron for assistance with mouse AGM dissection and pre-HSC sorting; and Fuchou Tang and Shuhui Bian for assistance with multi-omic experiments and analysis. We thank Qiu Wu from the Klein lab for constructive comments on our manuscript. F.D.C. is funded by grants R01HL128850, RC2DK131963, R01HL158192, the Edward P Evans Foundation, and an Alex Lemonade Crazy 8 award. S.-W.W. acknowledges support from the Westlake High-Performance Computing Center. S.B. acknowledges support from EMBO (ALTF 798-2018), the Wellcome Trust (grant number 215920/Z/19/Z), and the NIH (grant number 1K99HL164969). Illustrative figures are created with BioRender.

AUTHOR CONTRIBUTIONS

L.L., S.-W.W., and F.D.C. conceived the project, designed the study, and analyzed the data. L.L., S.-W.W., and F.D.C. wrote the manuscript with help from all other authors (especially S.E.M.). L.L. generated Cas9-TdT, Cas9-TdT-gRNAs and Cas9-TdT-gRNAs-TA mESCs, and Cas9-TdT, Cas9-TdT-gRNAs mouse lines. S.B. and B.L. produced Tigre- and Rosa26 target-array mouse lines. L.L. developed the Camellia-seq method. L.L. performed the mESCs, animal, and sequencing experiments with help from Q.Y. (mouse sample collection and cell sorting), K.A. (mouse sample collection), Y.J. (CARLIN library generation), X.L. (cell sorting and qPCR), S.B. (mouse AGM dissection), and M.F. (IP injection in mouse neonates). S.-W.W. conceived and developed the computational methods for analyzing DARLIN and Camellia-seq data, wrote the software packages, and performed all the statistical analyses. S.E.M. provided valuable feedback on data analysis and interpretation. A.M.K. supervised statistical analysis and lineage-tracing experimental design. F.D.C. and S.-W.W. supervised the study.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 23, 2023
Revised: July 11, 2023
Accepted: September 19, 2023
Published: October 17, 2023

REFERENCES

1. Kretzschmar, K., and Watt, F.M. (2012). Lineage tracing. Cell *148*, 33–45.
2. Bałakier, H., and Pedersen, R.A. (1982). Allocation of cells to inner cell mass and trophectoderm lineages in preimplantation mouse embryos. Dev. Biol. *90*, 352–362.
3. Snippert, H.J., van der Flier, L.G., Sato, T., van Es, J.H., van den Born, M., Kroon-Veenboer, C., Barker, N., Klein, A.M., van Rheenen, J., Simons, B.D., et al. (2010). Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. Cell *143*, 134–144.
4. Gerrits, A., Dykstra, B., Kalmykowa, O.J., Klauke, K., Verovskaya, E., Broekhuis, M.J.C., de Haan, G., and Bystrykh, L.V. (2010). Cellular barcoding tool for clonal analysis in the hematopoietic system. Blood *115*, 2610–2618.
5. Lu, R., Neff, N.F., Quake, S.R., and Weissman, I.L. (2011). Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nat. Biotechnol. *29*, 928–933.
6. Sun, J., Ramos, A., Chapman, B., Johnnidis, J.B., Le, L., Ho, Y.J., Klein, A., Hofmann, O., and Camargo, F.D. (2014). Clonal dynamics of native haematopoiesis. Nature *514*, 322–327.
7. Pei, W., Feyerabend, T.B., Rössler, J., Wang, X., Postrach, D., Busch, K., Rode, I., Klapproth, K., Dietlein, N., Quedenau, C., et al. (2017). Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. Nature *548*, 456–460.
8. Rodriguez-Fraticelli, A.E., Wolock, S.L., Weinreb, C.S., Panero, R., Patel, S.H., Jankovic, M., Sun, J., Calogero, R.A., Klein, A.M., and Camargo, F.D. (2018). Clonal analysis of lineage fate in native haematopoiesis. Nature *553*, 212–216.
9. Patel, S.H., Christodoulou, C., Weinreb, C., Yu, Q., da Rocha, E.L., Pepe-Mooney, B.J., Bowling, S., Li, L., Osorio, F.G., Daley, G.Q., et al. (2022). Lifelong multilineage contribution by embryonic-born blood progenitors. Nature *606*, 747–753.
10. Pei, W., Shang, F., Wang, X., Fanti, A.K., Greco, A., Busch, K., Klapproth, K., Zhang, Q., Quedenau, C., Sauer, S., et al. (2020). Resolving fates and single-cell transcriptomes of hematopoietic stem cell clones by PolyloxExpress barcoding. Cell Stem Cell *27*, 383–395.e8.
11. McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science *353*, aaf7907.
12. Alemany, A., Florescu, M., Baron, C.S., Peterson-Maduro, J., and van Oudenaarden, A. (2018). Whole-organism clone tracing using single-cell sequencing. Nature *556*, 108–112.
13. Raj, B., Wagner, D.E., McKenna, A., Pandey, S., Klein, A.M., Shendure, J., Gagnon, J.A., and Schier, A.F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nat. Biotechnol. *36*, 442–450.
14. Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., and Junker, J.P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. Nat. Biotechnol. *36*, 469–473.
15. Chan, M.M., Smith, Z.D., Grosswendt, S., Kretzmer, H., Norman, T.M., Adamson, B., Jost, M., Quinn, J.J., Yang, D., Jones, M.G., et al. (2019). Molecular recording of mammalian embryogenesis. Nature *570*, 77–82.
16. Kalhor, R., Kalhor, K., Mejia, L., Leeper, K., Graveline, A., Mali, P., and Church, G.M. (2018). Developmental barcoding of whole mouse via homing CRISPR. Science *361*, eaat9804.
17. Yang, D., Jones, M.G., Naranjo, S., Rideout, W.M., 3rd, Min, K.H.J., Ho, R., Wu, W., Replogle, J.M., Page, J.L., Quinn, J.J., et al. (2022). Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. Cell *185*, 1905–1923.e25.
18. Quinn, J.J., Jones, M.G., Okimoto, R.A., Nanjo, S., Chan, M.M., Yosef, N., Bivona, T.G., and Weissman, J.S. (2021). Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. Science *371*, eabc1944.
19. Bowling, S., Sritharan, D., Osorio, F.G., Nguyen, M., Cheung, P., Rodriguez-Fraticelli, A., Patel, S., Yuan, W.-C., Fujiwara, Y., Li, B.E., et al. (2020). An engineered CRISPR-Cas9 mouse line for simultaneous readout of lineage histories and gene expression profiles in single cells. Cell *181*, 1410–1422.e27.
20. Biddy, B.A., Kong, W., Kamimoto, K., Guo, C., Waye, S.E., Sun, T., and Morris, S.A. (2018). Single-cell mapping of lineage and identity in direct reprogramming. Nature *564*, 219–224.
21. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science *367*, eaaw3381.
22. Rodriguez-Fraticelli, A.E., Weinreb, C., Wang, S.W., Migueles, R.P., Jankovic, M., Usart, M., Klein, A.M., Lowell, S., and Camargo, F.D. (2020). Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. Nature *583*, 585–589.
23. Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., et al. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296–309.
24. Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.
25. Kelsey, G., Stegle, O., and Reik, W. (2017). Single-cell epigenomics: recording the past and predicting the future. Science *358*, 69–75.
26. Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.
27. Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., et al. (2014). Immunogenetics. Chromatin state dynamics during blood formation. Science *345*, 943–949.
28. Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.
29. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al. (2020). Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell *183*, 1103–1116.e20.
30. Lee, J., Hyeon, D.Y., and Hwang, D. (2020). Single-cell multiomics: technologies and data analysis methods. Exp. Mol. Med. *52*, 1428–1442.

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

32. Jindal, K., Adil, M.T., Yamaguchi, N., Wang, H.C., Yang, X., Kamimoto, K., Rivera-Gonzalez, G.C., and Morris, S.A. (2022). Multiomic single-cell lineage tracing to dissect fate-specific gene regulatory programs. bioRxiv.

33. Bian, S., Hou, Y., Zhou, X., Li, X., Yong, J., Wang, Y., Wang, W., Yan, J., Hu, B., Guo, H., et al. (2018). Single-cell multiomics sequencing and analyses of human colorectal cancer. Science *362*, 1060–1063.

34. Mimitou, E.P., Lareau, C.A., Chen, K.Y., Zorzetto-Fernandes, A.L., Hao, Y., Takeshima, Y., Luo, W., Huang, T.S., Yeung, B.Z., Papalexi, E., et al. (2021). Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat. Biotechnol. *39*, 1246–1258.

35. Boulé, J.B., Rougeon, F., and Papanicolaou, C. (2000). Comparison of the two murine terminal [corrected] deoxynucleotidyltransferase terminal isoforms. A 20-amino acid insertion in the highly conserved carboxyl-terminal region modifies the thermosensitivity but not the catalytic activity. J. Biol. Chem. *275*, 28984–28988.

36. Boulé, J.B., Rougeon, F., and Papanicolaou, C. (2001). Terminal deoxynucleotidyl transferase indiscriminately incorporates ribonucleotides and deoxyribonucleotides. J. Biol. Chem. *276*, 31388–31393.

37. Roychoudhury, R., Jay, E., and Wu, R. (1976). Terminal labeling and addition of homopolymer tracts to duplex DNA fragments by terminal deoxynucleotidyl transferase. Nucleic Acids Res. *3*, 101–116.

38. Loveless, T.B., Grotts, J.H., Schechter, M.W., Forouzmand, E., Carlson, C.K., Agahi, B.S., Liang, G., Ficht, M., Liu, B., Xie, X., et al. (2021). Lineage tracing and analog recording in mammalian cells by single-site DNA writing. Nat. Chem. Biol. *17*, 739–747.

39. Beard, C., Hochedlinger, K., Plath, K., Wutz, A., and Jaenisch, R. (2006). Efficient method to generate single-copy transgenic mice by site-specific integration in embryonic stem cells. Genesis *44*, 23–28.

40. Bunge, J., Willis, A., and Walsh, F. (2014). Estimating the number of species in microbial diversity studies. Annu. Rev. Stat. Appl. *1*, 427–445.

41. Chao, A. (1984). Nonparametric estimation of the number of classes in a population. Scand. Stat. Theory Appl. *11*, 265–270.

42. Laurenti, E., and Göttgens, B. (2018). From haematopoietic stem cells to complex differentiation landscapes. Nature *553*, 418–426.

43. Nishikii, H., Kanazawa, Y., Umemoto, T., Goltsev, Y., Matsuzaki, Y., Matsushita, K., Yamato, M., Nolan, G.P., Negrin, R., and Chiba, S. (2015). Unipotent megakaryopoietic pathway bridging hematopoietic stem cells and mature megakaryocytes. Stem Cells *33*, 2196–2207.

44. Roch, A., Trachsel, V., and Lutolf, M.P. (2015). Brief report: single-cell analysis reveals cell division-independent emergence of megakaryocytes from phenotypic hematopoietic stem cells. Stem Cells *33*, 3152–3157.

45. Carrelha, J., Meng, Y., Kettyle, L.M., Luis, T.C., Norfo, R., Alcolea, V., Boukarabila, H., Grasso, F., Gambardella, A., Grover, A., et al. (2018). Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells. Nature *554*, 106–111.

46. Morcos, M.N.F., Li, C., Munz, C.M., Greco, A., Dressel, N., Reinhardt, S., Sameith, K., Dahl, A., Becker, N.B., Roers, A., et al. (2022). Fate mapping of hematopoietic stem cells reveals two pathways of native thrombopoiesis. Nat. Commun. *13*, 4504.

47. Mikkola, H.K.A., and Orkin, S.H. (2006). The journey of developing hematopoietic stem cells. Development *133*, 3733–3744.

48. Kristiansen, T.A., Zhang, Q., Vergani, S., Boldrin, E., Krausse, N., André, O., Nordenfelt, P., Sigvardsson, M., Bryder, D., Ungerbäck, J., et al. (2022). Developmental cues license megakaryocyte priming in murine hematopoietic stem cells. Blood Adv. *6*, 6228–6241.

49. Wang, S.W., Herriges, M.J., Hurley, K., Kotton, D.N., and Klein, A.M. (2022). CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. Nat. Biotechnol. *40*, 1066–1074.

50. Wright, D.E., Wagers, A.J., Gulati, A.P., Johnson, F.L., and Weissman, I.L. (2001). Physiological migration of hematopoietic stem and progenitor cells. Science *294*, 1933–1936.

51. Abkowitz, J.L., Robinson, A.E., Kale, S., Long, M.W., and Chen, J. (2003). Mobilization of hematopoietic stem cells during homeostasis and after cytokine exposure. Blood *102*, 1249–1253.

52. Massberg, S., Schaerli, P., Knezevic-Maramica, I., Köllnberger, M., Tubo, N., Moseman, E.A., Huff, I.V., Junt, T., Wagers, A.J., Mazo, I.B., et al. (2007). Immunosurveillance by hematopoietic progenitor cells trafficking through blood, lymph, and peripheral tissues. Cell *131*, 994–1008.

53. Takihara, Y., Higaki, T., Yokomizo, T., Umemoto, T., Ariyoshi, K., Hashimoto, M., Sezaki, M., Takizawa, H., Inoue, T., Suda, T., et al. (2022). Bone marrow imaging reveals the migration dynamics of neonatal hematopoietic stem cells. Commun. Biol. *5*, 776.

54. Li, Y., Kong, W., Yang, W., Patel, R.M., Casey, E.B., Okeyo-Owuor, T., White, J.M., Porter, S.N., Morris, S.A., and Magee, J.A. (2020). Single-cell analysis of neonatal HSC Ontogeny Reveals gradual and uncoordinated transcriptional reprogramming that begins before birth. Cell Stem Cell *27*, 732–747.e7.

55. Li, L., Li, L., Li, Q., Liu, X., Ma, X., Yong, J., Gao, S., Wu, X., Wei, Y., Wang, X., et al. (2021). Dissecting the epigenomic dynamics of human fetal germ cell development at single-cell resolution. Cell Res. *31*, 463–477.

56. Clark, S.J., Argelaguet, R., Kapourani, C.A., Stubbs, T.M., Lee, H.J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J.C., et al. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat. Commun. *9*, 781.

57. Argelaguet, R., Clark, S.J., Mohammed, H., Stapel, L.C., Krueger, C., Kapourani, C.A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C.W., et al. (2019). Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature *576*, 487–491.

58. Fan, X., Lu, P., Wang, H., Bian, S., Wu, X., Zhang, Y., Liu, Y., Fu, D., Wen, L., Hao, J., et al. (2022). Integrated single-cell multiomics analysis reveals novel candidate markers for prognosis in human pancreatic ductal adenocarcinoma. Cell Discov. *8*, 13.

59. Li, L., Dong, J., Yan, L., Yong, J., Liu, X., Hu, Y., Fan, X., Wu, X., Guo, H., Wang, X., et al. (2017). Single-cell RNA-Seq analysis maps development of human germline cells and gonadal niche interactions. Cell Stem Cell *20*, 891–892.

60. Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. *21*, 1160–1167.

61. Kelly, T.K., Liu, Y., Lay, F.D., Liang, G., Berman, B.P., and Jones, P.A. (2012). Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. Genome Res. *22*, 2497–2506.

62. Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat. Methods *11*, 817–820.

63. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: uniform manifold approximation and projection. J. Open Source Softw. *3*, 861.

64. Choi, J., Chen, W., Minkina, A., Chardon, F.M., Suiter, C.C., Regalado, S.G., Domcke, S., Hamazaki, N., Lee, C., Martin, B., et al. (2022). A time-resolved, multi-symbol molecular recorder via sequential genome editing. Nature *608*, 98–107.

65. Pettitt, S.J., Liang, Q., Rairdan, X.Y., Moran, J.L., Prosser, H.M., Beier, D.R., Lloyd, K.C., Bradley, A., and Skarnes, W.C. (2009). Agouti C57BL/6N embryonic stem cells for mouse genetic resources. Nat Methods *6*, 493–495.

66. Buchholz, F., Angrand, P.O., and Stewart, A.F. (1998). Improved properties of FLP recombinase evolved by cycling mutagenesis. Nat Biotechnol *16*, 657–662.

67. Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

68. Zhou, F., Li, X., Wang, W., Zhu, P., Zhou, J., He, W., Ding, M., Xiong, F., Zheng, X., Li, Z., et al. (2016). Tracing haematopoietic stem cell formation at single-cell resolution. Nature *533*, 487–492.

69. Baron, C.S., Kester, L., Klaus, A., Boisset, J.C., Thambyrajah, R., Yvernogeau, L., Kouskoff, V., Lacaud, G., van Oudenaarden, A., and Robin, C. (2018). Single-cell transcriptomics reveal the dynamic of haematopoietic stem cell production in the aorta. Nat. Commun. *9*, 2517.

70. Kremer, L.P.M., Küchenhoff, L., Cerrizuela, S., Martin-Villalba, A., and Anders, S. (2022). Analyzing single-cell bisulfite sequencing data with scbs. bioRxiv.. https://doi.org/10.1101/2022.06.15.496318.

71. Sfeir, A., and Symington, L.S. (2015). Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? Trends Biochem. Sci. *40*, 701–714.

72. Stinson, B.M., Moreno, A.T., Walter, J.C., and Loparo, J.J. (2020). A mechanism to minimize errors during non-homologous end joining. Mol. Cell *77*, 1080–1091.e8.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
FC Receptor Block	eBioscience	Cat# 156604
biotin-conjugated lineage antibody cocktail	Miltenyi Biotec	Cat# 130-090-858
Streptavidin eFluor® 450	eBioscience	Cat# 48-4317-82
Streptavidin PE-Cy7	eBioscience	Cat# 25-4317-82; RRID: AB_10116480
APC anti-mouse CD117 (c-kit) [ACK2]	BioLegend	Cat# 135108; RRID: AB_2028407
PE anti-mouse Ly-6A/E (Sca-1) [D7]	eBioscience	Cat# 12-5981-83; RRID: AB_466087
PE-Cy7 anti-mouse Ly-6A/E (Sca-1) [D7]	eBioscience	Cat# 25-5981-82; RRID: AB_469669
APC-Cy7 anti-mouse CD48 [HM48-1]	BD Pharmingen	Cat# 561242
PE-Cy5 anti-mouse CD150 [TC15-12F12.2]	BioLegend	Cat# 115912; RRID: AB_493598
BV605 anti-mouse CD41 [MWReg30]	BioLegend	Cat# 133921; RRID: AB_2563933
APC anti-mouse/human CD45R/B220 [RA3-6B2]	BioLegend	Cat# 17-0452-83
PE anti-mouse Ly-6G [1A8]	BioLegend	Cat# 127608; RRID: AB_1186099
APC-Cy7 anti-mouse CD11b (Mac1) [M1/70]	BD Pharmingen	Cat# 557657
PE-Cyanine5 anti-mouse Ter-119 [TER-119]	eBioscience	Cat# 15-5921-83; RRID: AB_468811
Brilliant Violet 605™ anti-mouse CD45 Antibody [30-F11]	BioLegend	Cat# 103140; RRID: AB_2562342
PerCP/Cy5.5 anti-mouse CD31 [MEC13.3]	BioLegend	Cat# 102522; RRID: AB_2566761
Anti-SSEA-1 Mouse Monoclonal Antibody (Biotin) [clone: MC-480]	BioLegend	Cat# 125604; RRID: AB_1089194
DAPI (4',6-Diamidino-2-Phenylindole, Dihydrochloride)	Invitrogen	Cat# D1306; RRID: AB_2629482
Chemicals, Peptides, and Recombinant Proteins		
Doxycycline	Sigma	Cat# D9891
Knock-Out DMEM	GIBCO	Cat# 10829018
ESC US Qualified FBS	GIBCO	Cat# 16141079
LIF	EMD Millipore	Cat# ESG1107
Chiron	Sigma	Cat# SML1046
Mirdametinib	Selleck Chemicals	Cat# S1036
Non-essential amino acids	GIBCO	Cat# 11140-050
L-Glutamine	GIBCO	Cat# 25030164
Penicillin-Streptomycin	GIBCO	Cat# 15140-163
MEM NEAA	GIBCO	Cat# 11140050
β-Mercaptoethanol	GIBCO	Cat# 21985023
Mitomycin C	Santa-Cruz	Cat# CAS 50-07-7
Hygromycin	GIBCO	Cat# 10687010
Puromycin	Sigma-Aldrich	Cat# P9620
fetal bovine serum (FBS)	Gibco	Cat# A3160401
Trypsin-EDTA (0.05%)	GIBCO	Cat# 25300120
Collagenase	Sigma-Aldrich	Cat# C0130
Trizol	Invitrogen	Cat# 15596018
Qiagen protease	Qiagen	Cat# 19155
Superscript III Reverse Transcriptase	Invitrogen	Cat# 18080-044

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Q5 polymerase	New England Biolabs (NEB)	Cat# M0491
KAPA HiFi HotStart ReadyMix	Roche Applied Science	Cat# 07958935001
dNTPs	New England Biolabs (NEB)	Cat# N0446
SuperScript® II Reverse Transcriptase	Thermo Scientific	Cat# 18064014
GpC Methyltransferase (M.CviPI)	New England Biolabs (NEB)	Cat# M0227L
PhiX Control v3	Illumina	Cat# FC-110-3001
Critical Commercial Assays		
Nebuilder HiFi DNA Assembly	New England Biolabs (NEB)	Cat# E5520S
Endotoxin-free plasmid maxiprep kit	Takara	Cat# 740424.10
Mouse Nucleofector Kit	Lonza	Cat# VPH-1001
MACS Separation Columns	Miltenyi Biotec	Cat# 130-042-401
AMPure XP beads	Beckman Coulter	Cat# A63881
SPRIselect reagent	Beckman Coulter	Cat# B23318
Dynabeads Myone Carboxylic Acid	Invitrogen	Cat# 65011
Chromium Next GEM Single Cell 3′ Kit v3.1	10X Genomics	Cat# PN-1000268
Dual Index Kit TT Set A	10x Genomics	Cat# PN-1000215
EZ-96 DNA Methylation-Direct MagPrep	Zymo Research	Cat# D5044
Klenow Fragment (3′-5′ exo-)	New England Biolabs (NEB)	Cat# M0212L
KAPA Library Quantification Kit	Kappa Biosystems	Cat# KK4835
MiSeq Reagent Kit v2 (500 cycles)	Illumina	Cat# MS-102-2003
Deposited data		
Raw and processed data	This paper	GEO: GSE222486
Experimental Models: Cell Lines		
DR4-MEFs	ATCC	SCRC-1045
KH2 mESCs	Beard et al.[39]	N/A
JM8A3 mESCs	Pettitt et al.[65]	N/A
Cas9–TdT mESCs	This paper	N/A
Cas9-TdT-gRNAs mESCs	This paper	N/A
Tigre-target-array (TA) mESCs	This paper	N/A
Rosa26-target-array (RA) mESCs	This paper	N/A
Cas9-TdT-gRNAs-TA mESCs	This paper	N/A
Experimental Models: Organisms/Strains		
Mouse: B6;129S4-Gt(ROSA) $26Sor^{tm1(rtTA*M2)Jae}$ $Col1a1^{tm1(tetO-cas9)Sho}$/J	Stuart H. Orkin	Cat# JAX:029415; RRID: IMSR_JAX:029415
Col1a1-target-array mouse: B6;129S4-$Col1a1^{tm4(CAG-EGFP)Fcam}$/Mmjax	Fernando Camargo	MMRRC Strain #067061-JAX; RRID: MMRRC_067061-JAX
Cas9-TdT mouse: B6;129S4-Gt(ROSA) $26Sor^{tm1(rtTA*M2)Jae}$ $Col1a1^{tm1(Cas9-TdT)}$	This paper	N/A
Cas9-TdT-gRNA mouse: B6;129S4-Gt(ROSA)$26Sor^{tm1(rtTA*M2)Jae}$ $Col1a1^{tm5(tetO-Cas9/Dntt*)Fcam}$/J	This paper	Cat# JAX:038749
CA-TA-RA mouse: B6;129S4-Gt(ROSA) $26Sor^{tm1(UBC-GFP)Fcam}$ $Igs7^{em1(CAG-mCherry)Fcam}$ $Col1a1^{tm4(CAG-EGFP)Fcam}$/J	This paper	Cat# JAX:038750
Oligonucleotides		
Primers, see Table S1	IDT	N/A
Recombinant DNA		
Plasmid: pCAG-Cas9	Addgene	Cat# 48138; RRID: Addgene_48138

(*Continued on next page*)

Please cite this article in press as: Li et al., A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells, Cell (2023), https://doi.org/10.1016/j.cell.2023.09.019

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Plasmid: pCAG-Cas9-TdT	This paper	N/A
Plasmid: pCAGGS-FLPe-puro	Buchholz et al.[66]	N/A
Plasmid: pBS31 targeting vector	Beard et al.[39]	N/A
Plasmid: pBS31-Cas9-TdT	This paper	N/A
Plasmid: pBS31-Cas9-TdT-gRNAs	This paper	N/A
Plasmid: px459	Addgene	Cat# 48139; RRID: Addgene_48139
Plasmid: Tigre-targeting vector	Addgene	Cat# 92142; RRID: Addgene_92142
Plasmid: pCAG-mCherry-Tigre-target-array	This paper	N/A
Plasmid: Rosa26-targeting vector	Addgene	Cat# 74286; RRID: Addgene_74286
Plasmid: pUBC-GFP-Rosa26-target-array	This paper	N/A
Software and Algorithms		
BioRender	BioRender	https://biorender.com/
FlowJo	FlowJo, LLC	https://www.flowjo.com/solutions/flowjo/downloads
CoSpar v0.3.0	Wang et.al.[49]	https://cospar.readthedocs.io/
CARLIN pipeline	Bowling et.al.[19]	https://gitlab.com/hormozlab/carlin
snakemake_DARLIN	This study	https://github.com/ShouWenWang-Lab/snakemake_DARLIN
MosaicLineage	This study	https://github.com/ShouWenWang-Lab/MosaicLineage
Bismark v.0.23.0	Krueger and Andrews[67]	https://www.bioinformatics.babraham.ac.uk/projects/bismark/
CellRanger v.7.0.0	10X Chromium	https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/installation

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Fernando Camargo (fernando.camargo@childrens.harvard.edu).

Materials availability

Plasmids generated by this study are available upon request and will be deposited to Addgene. Mouse lines generated by this study are available upon request and will be deposited to The Jackson Laboratory.

Data and code availability

The accession number for the sequencing data reported in this paper is NCBI GEO: GSE222486. The snakemake pipeline for allele preprocessing is available at https://github.com/ShouWenWang-Lab/snakemake_DARLIN while the companion python package for downstream analysis is available at https://github.com/ShouWenWang-Lab/MosaicLineage.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Mice

Six transgenic mouse lines were used in this study: iCas9,[19] cCARLIN,[19] Cas9–TdT, Cas9–TdT-gRNAs, Tigre-target-array (TA), and Rosa26-target-array (RA). Among them, Cas9–TdT, Cas9-TdT-gRNAs, Tigre-target-array, and Rosa26-target-array were generated in this study. Mouse generation is detailed in method details. For timed pregnancy experiments, we housed female and male mice together overnight and checked the female the following morning for a vaginal plug. The detection of a plug marked embryonic day 0.5 (E0.5). All animal procedures were approved by the Boston Children's Hospital Institutional Animal Care and Use Committee.

ESC lines

We generated five mouse ESC (mESC) lines in this study: Cas9–TdT KH2, Cas9–TdT-gRNAs KH2, Tigre-target-array (TA) JM8A3, Rosa26-target-array (RA) JM8A3 and Cas9-TdT-gRNAs-TA KH2 mESC line. See method details for more information.

METHOD DETAILS

Cas9–TdT Fusion Protein Design

The Cas9–TdT fusion protein was designed using the coding sequence of SpCas9, which contains an N-terminal 3×FLAG peptide followed by the SV40-NLS and a C-terminal nucleoplasmin NLS, and the coding sequence of the d138 TdT mutant, which contains an N-terminal 6×His tag. The fusion protein sequence was generated by replacing the stop codon of the SpCas9 coding sequence with a linker sequence (GGGGSGGGGSGGGGS) followed by the entire d138 TdT mutant coding sequence. The corresponding 3′ UTR for this coding sequence includes a WPRE and β-globin poly(A) signal sequence. A DNA sequence containing the Cas9–TdT coding sequence and its corresponding 3′ UTR was synthesized by Vectorbuilder and cloned into the pCAG backbone.

Hydrodynamic Tail Vein Injection of Cas9–TdT or Cas9 Plasmid into CARLIN Mice

pCAG-Cas9 (Addgene Plasmid #48138) and pCAG-Cas9–TdT plasmids were prepared by endotoxin-free plasmid maxiprep kit (Takara Cat.#740424.10) and adjusted to a concentration of 10 μg/mL in DPBS (Life Technologies, Cat.#14190250). 20 μg (2 mL) pCAG-Cas9 or pCAG-Cas9–TdT plasmid was injected into CARLIN mice (*n* = 2 for each condition) through hydrodynamic tail vein injections. After one week, livers were collected from all four mice and homogenized in Trizol (Life Technologies, Cat.#15596018) with a tissue homogenizer (IKA, 0003737001). Total RNA was subsequently purified with Trizol and quantified using a Nanodrop One spectrophotometer (Thermo Scientific, Cat.#13-400-518).

Generation of Cas9–TdT-related ESC Lines

Three Cas9–TdT-related ESC lines were generated in the course of this study: Cas9–TdT KH2 ESC line, Cas9–TdT-gRNAs KH2 ESC line, and Cas9–TdT-gRNAs-TA KH2 ESC line. We used the KH2 ESC line that carries a donor FRT site in the *Col1a1* locus and M2-rtTA in the *Rosa26* locus.[39] KH2 ESCs were cultured on DR4-MEFs feeder layer (ATCC, SCRC-1045) and maintained in the following medium, hereafter referred to as Standard ESC Medium: Knock-Out DMEM (GIBCO, 10829018), 15% ESC US Qualified Fetal Bovine Serum (FBS) (GIBCO, 16141079), 100 U/mL Penicillin-Streptomycin (Pen-Strep) (GIBCO, 15140-163), 2 mM L-Glutamine (GIBCO, 25-030-081), MEM Non-Essential Amino Acids (NEAA) (GIBCO, 11140050), 0.1 mM β-Mercaptoethanol (GIBCO, 21985023) and 10 ng/mL Leukemia Inhibitory Factor (LIF) (EMD Millipore, ESG1107).

To generate the Cas9–TdT KH2 ESC line, we cloned Cas9–TdT into pBS31 targeting vector[39] which contains a FRT site compatible with the Flippase to generate the pBS31-Cas9–TdT plasmid, where Cas9–TdT is regulated by the Tet-On doxycycline inducible promoter and contains a WPRE within its 3′ UTR (Figure 2A, top-most row). We next integrated the Cas9–TdT-WPRE sequence into the *Col1a1* locus through expression of a flippase recombinase construct (pCAGGS-FLPe-puro).[39] Approximately ten million KH2 ESCs were nucleofected with 17 μg of the pBS31-Cas9-TdT vector and 8 μg of pCAGGS-FLPe-puro using theAmaxa Nucleofector II (setting A-23) with the Mouse Nucleofector Kit (Lonza VPH-1001). After one day, 140 mg/mL hygromycin (GIBCO, 10687010) was used to select positive clones. After ten days of selection, individual clones were picked and genotyped using Cas9–TdT primers and *Col1a1*-flanking primers to ensure the correct integration. We then measured protein expression of Cas9–TdT upon Dox induction for all positively genotyped clones by western-blot. The Cas9–TdT KH2 ESC clone with the highest protein expression was used to generate the mouse line.

For the Cas9–TdT-gRNAs KH2 ESC line, we cloned the element containing ten tandem gRNAs downstream of Cas9–TdT with a reverse expression direction to get pBS31-Cas9–TdT-gRNAs vector (Figure 3A, second row). Using the same protocol described above, we integrated Cas9–TdT-gRNAs into the *Col1a1* locus of KH2 ESCs. The selected Cas9-TdT-gRNAs KH2 ESC clone was used to generate the mouse line.

For the Cas9–TdT-gRNAs-TA KH2 ESC line, Cas9–TdT-gRNAs KH2 ESC line was subjected to another round of targeting to introduce mCherry-target array into the *Tigre* locus using homologous recombination (Figure 3A, 2nd and 4th row). The Tigre-target-array, made up of identical guide RNA target sequences to the original Col1a1 target array in a different order, was incorporated downstream of a WPRE into the 3' UTR region of a mCherry open reading frame, upstream of a SV40-poly(A) signal (construct synthesized by Genewiz). This construct was cloned into Tigre-targeting vector (Addgene #92142) downstream of the CAG promoter to generate the pCAG-mCherry-Tigre-target-array plasmid. 10 million Cas9-TdT-gRNAs KH2 ESCs were nucleofected using 25 μg linearized pCAG-mCherry-Tigre-target-array vector (via AvrII endonuclease, NEB, #R0174L) and 25 μg px459 (Addgene # 48139) that contains a Tigre-locus-targeting gRNA sequence (ACTGCCATAACACCTAACTT) to increase efficiency of homologous recombination. mCherry-positive ESCs were selected by FACS after three days and reseeded on MEFs to grow for another week. Single ESC clones were picked up and genotyped with primers specific for the *Tigre* locus and TA. The selected Cas9-TdT-gRNAs-TA KH2 ESC clone was used to study the editing kinetics of Cas9–TdT (Figure S1).

Generation of Tigre and Rosa26 target-array ESC lines

pCAG-mCherry-Tigre-target-array was linearized via digestion with AvrII and introduced into JM8A3 mESCs by nucleofection. Nucleofection was performed using 25 μg of linear pCAG-mCherry-Tigre-target-array plasmid together with 25 μg of px459

Please cite this article in press as: Li et al., A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells, Cell (2023), https://doi.org/10.1016/j.cell.2023.09.019

(Addgene # 48139) carrying Cas9 and Tigre-locus-targeting gRNA to facilitate homologous recombination. Cells were selected with puromycin for two days (2 μg/ml; Sigma-Aldrich #P9620) and cultured for a further 7 days before mCherry-positive ESC colonies were picked for further culture. Colonies were then screened both for correct integration of the Tigre-target-array via Sanger sequencing and for lack of Cas9 integration via PCR, to ensure the px459 plasmid was not stably integrated. JM8A3 mESCs were cultured in the Standard ESC Medium (described above) further supplemented with 30 mM Chiron (Sigma SML1046) and 10 mM Mirdametinib (Selleck Chemicals S1036).

The Rosa26-target-array, which contains the same ten gRNA target sites as the original Col1a1-target-array but in a different order, was incorporated into the 3′ UTR of a GFP expression construct, downstream of a WPRE element and upstream of a SV40-poly(A) signal sequence and 2xHS4 insulator elements, all under the control of a UBC promoter (construct synthesized by Genewiz). This construct was cloned into a Rosa26-targeting vector (Addgene #74286) in the opposite orientation to the Rosa26 promoter to generate the pUBC-GFP-Rosa26-target-array plasmid. The vector also contained a CAG-puro construct for mESC selection. The plasmid was linearized via digestion with SgrDI (Thermo Scientific, #ER2031), and then 25 μg of linearized plasmid was introduced into JM8A3 cells via nucleofection, as described above. Cells were selected with puromycin for nine days, after which GFP-positive colonies were picked for further culture, and colonies were screened for correct integration of the Rosa26-target-array by Sanger sequencing. Rosa26-target-array mESCs were cultured in identical conditions to those of the Tigre-target-Array mESCs.

Generation of Tigre and Rosa26 target-array mouse lines

One clone each of the Tigre and Rosa26 target-array ESCs were selected for injection into C57/BL6 blastocysts to form chimeras. Chimera generation was performed by the Gene Manipulation & Genome Editing Core at Boston Children's Hospital. High contribution male chimeras were selected for further breeding to C57/BL6 female mice to obtain F1 mice. These lines were crossed to each other and to the Col1a1-target-array mice[19] to generate triple-array (CA/TA/RA) homozygous mice.

Generation of Cas9–TdT Mouse Lines

The selected Cas9–TdT KH2 ESC clone was injected into C57/BL6 mouse embryos to generate mouse chimeras. High-contribution male chimeras were backcrossed into C57/BL6 mice to confirm germline transmission of the edited genome. Offsprings from F1 were crossed to Col1a1-target-array mice to generate the DARLIN-v0 mouse line. To generate the DARLIN mouse line, we repeated the above steps with Cas9–TdT-gRNAs KH2 ESC clone, and in the last step, we crossed the offsprings from F1 to mice homozygous for all three target arrays (CA, TA, and RA).

Administration of Doxycycline in Mice

To label adult mice, doxycycline (Sigma-Aldrich, D9891) was administered via drinking water for one week (2 mg/mL, supplemented with 10 mg/mL sucrose) and three intraperitoneal injections (50 μg/g) every other day during the same week. To label neonatal mice (P2-P6), Dox was administered via drinking water for four days (2 mg/mL supplemented with 10 mg/mL sucrose) and two intraperitoneal injections (50 μg/g). To label mouse embryos, 50 μg/g Dox was injected once into pregnant dams through a retro-orbital route.

Tissue Preparation

To sort pre-HSCs in AGM at E11.5, the AGM region was isolated as previously described,[9,68,69] and CD31 and cKit antibodies were injected into the aorta and incubated for 30 minutes to specifically label blood clusters. We dissociated the AGM tissues using 0.12% Collagenase (Sigma-Aldrich, C0130-500MG) at 37 °C for 15 minutes to get the cell suspension and filtered the cell suspension through a 40-μm strainer.

To prepare blood populations from fetal livers at E15.5, the livers were dissected and put in DPBS supplemented with 10% FBS (Gibco, A3160401) and 100 U/mL Pen-Strep. We gently pipetted twenty times to dissociate fetal livers, and then removed erythrocytes by incubating the material in RBC lysis buffer followed by collection of intact cells upon passing the lysate through a 40-μm strainer. To enrich for HSPCs, we conducted blood lineage depletion using a biotin-conjugated lineage antibody cocktail (Miltenyi Biotec, 130-090-858) followed by incubation with anti-biotin magnetic beads (Miltenyi Biotec, 130-090-858) and subsequent magnetic column depletion.

To isolate blood populations from adult bone marrow, different bones (spinal cord, skull, leg bone (including femur, tibia and fibula), and arm bone (including humerus, ulna and radius)) were dissected and crushed in DPBS supplemented with 2% FBS (Gibco, A3160401) and 100 U/mL Pen-Strep. Erythrocytes were then removed as described above. Blood lineage depletion was performed to enrich HSPCs or $Lin^{-}Sca1^{+}cKit^{+}$ cells (LSKs).

FACS

Cell suspensions were processed using FC Receptor Block (eBioscience, 48-0161-82), stained and sorted based on the following antibody panels: preHSC: CD31+cKit+Ssea1−; LT-HSC: Lin−cKit+Sca1+CD150+CD48−; HSC: Lin−cKit+Sca1+CD48−; MPP: Lin−cKit+Sca1+CD150−CD48+; LSK: Lin-cKit+Sca1+; LK: Lin-cKit+; MkP: Lin−cKit+Sca1−CD150+CD41+; Erythrocyte progenitor: Ter119+; Granulocyte: CD11b+Ly6G+; Monocyte: CD11b+Ly6G−; and B cell: B220+CD11b−Ly6G−. Lin- stands for Ter119-B220-CD19-CD3-CD11b-Gr1- cells. For AGM and fetal liver samples, an anti-CD45 antibody was also used to label blood cells. DAPI was added prior to sorting to exclude dead cells. Cell sorting was performed using a BD FACSAria II sorter.

Please cite this article in press as: Li et al., A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells, Cell (2023), https://doi.org/10.1016/j.cell.2023.09.019

Bulk Lineage Array Library Preparation

To generate lineage array libraries from tissues or cells, we used bulk RNA or DNA as starting material and a nested PCR approach to perform targeted amplification with Q5 High-Fidelity DNA Polymerase (New England Biosciences, M0491L). The primers used in the following protocols were included in the Table S1. We used 1.5X Ampure XP beads (Beckman Coulter, A63881) to purify the PCR product from each step once except the final indexing PCR product (0.8X, twice purification). The final indexing libraries were pooled and quantified (Kappa Biosystems, KK4835), and sequenced on an Illumina MiSeq using paired-end 500-cycle v2 kits (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).

To amplify *Col1a1* target array (CA) array–containing mRNAs from mouse livers (Figures 1G–1L), 500 ng of total RNA was reverse transcribed with SuperScrit III (Invitrogen, 18080093) and the CA-specific primer (RT_CA_10UMI). After cDNA purification, the first PCR reaction was performed for 15 cycles with the primers NGS_F and NGS_CA_R1. The resulting PCR product was purified, and then the second PCR reaction was performed for 15 cycles with the same NGS_F primer and a nested NGS_CA_R2 primer. The PCR product was further purified and used to perform an indexing PCR for 8 cycles using the primers including in this kit (New England Biosciences, E7500S).

To amplify the *Tigre* target array (TA) array from genomic DNA of ESCs (Figures S1A–S1D), $\sim 10^5$ cells were pelleted and resuspended in 20 μl of lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630 (Sigma-Aldrich, I8896), 0.5 mg/ml Qiagen Protease (Qiagen, 19155)). Then we performed a UMI-tagging reaction using the released genomic DNA, Q5 High-Fidelity DNA polymerase, and the RT_TA_10UMI primer.[11] Following purification, a targeted PCR (NGS_F and NGS_TA_R1,15 cycles) and a nested PCR (NGS_F and NGS_TA_R2,15 cycles) were performed. Finally, an indexing PCR was carried out (8 cycles; New England Biosciences, E7500S).

To prepare CA-, TA-, and RA-array libraries from the total RNA of blood cells, reverse transcription was performed with Superscript III and a mixture of three primers (RT_CA_12UMI, RT_TA_14UMI, and RT_RA_14UMI). After cDNA purification, we amplified each of the three target arrays individually by performing three separate PCR reactions for each sample. This was done due to the differences in expression between each of the three arrays, which would lead to preferential recovery of only the highly expressed TA array if the PCR were pooled. We first amplified the CA-, TA-, and RA-array libraries for 15 cycles using a 1/4th of the purified cDNA, the NGS_F primer, and either the NGS_CA_R1, NGS_TA_R1, or NGS_RA_R1 primer. The three PCR products were purified, and each was subject to a second PCR with 15 cycles, again using the NGS_F primer and either the NGS_CA_R2, NGS_TA_R2, or NGS_RA_R2 primer. Each of the resulting PCR products was further purified and used to perform an indexing PCR for 8 cycles using the primers included in this kit (New England Biosciences, E6609S).

Single-Cell Transcriptome and Lineage Array Library Preparation Based on 10X Genomics

We used the Chromium Next GEM Single Cell 3′ Kit v3.1 (10X Genomics, PN-1000268) and the Dual Index Kit TT Set A (10X Genomics, PN-1000215) to generate the single-cell, whole-transcriptome libraries, according to protocols provided by the manufacturer (https://cdn.10xgenomics.com/image/upload/v1668017706/support-documents/CG000315_ChromiumNextGEMSingleCell3-_GeneExpression_v3.1_DualIndex__RevE.pdf). The libraries were sequenced on an Illumina NovaSeq 6000 by Novogene with the paired-end 150 bp kit (Read 1: 28 cycles; i7 Index: 10 cycles; i5 Index: 10 cycles; Read 2: 90 cycles).

In parallel, we used KAPA HiFi HotStart ReadyMix (Roche Applied Science, 07958935001) to amplify single-cell CA/TA/RA libraries, and 1.5X SPRI SELECT REAGENT beads (Beckman Coulter, B23318) to purify the PCR product from each step except the final indexing product where we used 0.8X beads. Following the cleanup of cDNA (Step 2.2 of the above-referenced manufacturer's protocol), we amplified the triple target arrays separately as described in the "Bulk Lineage Array Library Preparation" section, with the following differences: 1.) Each of the first PCR reactions used 5–10 ng of purified cDNA. 2.) Each of the three first and second PCR steps was performed for 10 cycles and used the P5-PR1 primer instead of the NGS_F primer. Finally, the indexing PCR was conducted using the primers included in this kit (8 cycles; New England Biosciences, E6609S). 10X single-cell lineage array libraries were quantified using KAPA Library Quantification Kit and sequenced on an Illumina MiSeq using paired-end 500-cycle v2 kits (Read 1: 28 cycles; i7 Index: 8 cycles; Read 2: 350 cycles; Illumina, MS-102-2003) with 10% PhiX sequencing control v3 (Illumina, FC-110-3001).

Single-Cell Camellia-seq Library Preparation

Individual cells were directly sorted into 96-well plates containing the mild lysis buffer and a methyltransferase reaction mixture and incubated for 30 minutes at 37 °C. Dynabeads Myone Carboxylic Acid (Invitrogen, 65011) were used to capture nuclei and the supernatant containing released RNA was transferred to a separate 96-well plate. The Dynabeads containing genomic DNA were used to carry out single-cell bisulfite sequencing (scBS-seq)[62] to profile DNA methylation and chromatin accessibility. The RNA part was processed with a modified scSTRT-seq[59,60] to profile the whole transcriptome and cellular lineages. CA/TA/RA libraries were amplified from the cDNA using primers listed in Table S1.

QUANTIFICATION AND STATISTICAL ANALYSIS

Computational analysis overview

Analyses of lineage-tracing datasets generated using the DARLIN-v0 and DARLIN mice were performed by first identifying the complete set of distinct editing events (referred to as alleles or clonal barcodes) from the reads containing target array sequences, and then further estimating which of these alleles were likely to have been generated only once in the experiment (referred to as rare alleles) and therefore represent clonally related cells. For the scRNA-seq experiments (Figures 4, 6, 7, S4, S6, and S7), all three target-array loci were considered simultaneously to assign individual cells to their respective clones, while for the bulk RNA-seq experiments (Figures 5 and S5) only a single target array locus was used. Finally, these clonal assignments were used to perform the reported analyses, which include the calculation of clonal coupling scores between different cell types and/or tissues, the shared clone fraction for cell types in different tissues, fate-bias prediction for HSPCs, and evaluation of clonal memory using gene expression, DNA methylation, and chromatin accessibility (Figure S3A). We constructed an allele bank with a large collection of alleles to help to identify rare alleles (Figure S3B).

Allele preprocessing

The CARLIN pipeline (https://gitlab.com/hormozlab/carlin) was developed previously for the *Col1a1* target array (CA) in the Cas9/CARLIN mouse.[19] To call alleles from bulk or 10X single-cell target array sequences in our analysis, we updated the CARLIN workflow with minor modifications: 1) we expanded it to include the two additional target arrays (TA and RA) present in DARLIN mice by adding corresponding configuration files to the package; 2) we retained alleles supported by at least 3 reads per UMI or cell barcode, unless otherwise stated. The updated pipeline can be found at https://github.com/ShouWenWang-Lab/Custom_CARLIN. To call alleles from single-cell target array sequences from Camellia-seq, we developed a pipeline inspired by the original CARLIN pipeline with additional consideration for the sequencing errors from heterogeneous library size across cells, where the read cutoff was set to be 5 (https://github.com/ShouWenWang-Lab/snakemake_DARLIN/blob/master/QC/single_cell_CARLIN-scCamellia.ipynb).

We further developed a snakemake pipeline snakemake_DARLIN (https://github.com/ShouWenWang-Lab/snakemake_DARLIN) to automate the generation of alleles associated with each UMI or cell barcode from fastq files. This tool enables the processing of multiple samples in parallel. Finally, because the CARLIN pipeline was written in MatLab, we have developed a companion Python package MosaicLineage (https://github.com/ShouWenWang-Lab/MosaicLineage) so that these and any further generated DARLIN lineage data can be more easily analyzed using other single-cell analysis packages developed in Python.

Clone identification: theory

Suppose that M_0 distinct lineage barcodes have been detected, with each lineage barcode having a particular generation probability. Many of these barcodes will correspond to individual clones of cells, while others with a greater generation probability will independently occur within multiple clones (i.e., due to barcode homoplasy). Among the M_0 total barcodes, we wish to identify M clonal barcodes whose generation probability $\rho \leq \rho_*$, such that the false discovery rate (FDR) α within these M barcodes is below an appropriate threshold. Here, the FDR is defined as the fraction of the M ($\leq M_0$) selected barcodes that erroneously label more than one clone. Below, we calculate the FDR associated with a given ρ_*.

Assume that the total number of clones that are associated with our M barcodes is M_c ($\geq M$). In order to label each of the M_c clones, we sample M_c barcodes with replacement in a linear order out of the entire pool of M_T possible barcodes, and the probability associated with j-th sampled barcode is ρ_j ($1 \leq j \leq M_c$). Note that $\rho_j \leq \rho_*$ due to our filtering. When ρ_* is sufficiently small, and $M \ll M_T$, each newly sampled barcode is very likely different from previously sampled barcodes. Therefore, at j-th sampling, the probability of sampling a barcode that has already been sampled is approximately $\sum_{k=1}^{j-1}\rho_k$. Therefore, the average number of barcodes that erroneously label more than one clone is

$$m = \langle \sum_{j=1}^{M_c}\left(\sum_{k=1}^{j-1}\rho_k\right)\Bigg|\rho_k \leq \rho_*\rangle = \sum_{j=1}^{M_c}\sum_{k=1}^{j-1}\langle\rho_k|\rho_k \leq \rho_*\rangle = \frac{1}{2}M_c(M_c-1)\langle\rho|\rho \leq \rho_*\rangle.$$

Here, $\langle\rho|\rho \leq \rho_*\rangle$ is the average barcode generation probability below the cutoff ρ_*. Therefore, in order to satisfy FDR $m/M \leq \alpha$, we have

$$\langle\rho|\rho \leq \rho_*\rangle \leq \frac{2\alpha M}{M_c(M_c-1)} \approx \frac{2\alpha}{M-1},$$

where we have applied the approximation that $M_c \approx (1+\alpha)M \approx M$ when $\alpha \ll 1$, which is the regime of interest. $\langle\rho|\rho \leq \rho_*\rangle$ can be calculated from the empirical allele bank at a given cutoff ρ_*; M can be obtained from the observed data after removing barcodes with generation probability above ρ_*; and the FDR cutoff α can be tailored for each study.

Our above formulation deals with the general problem of clone identification when the barcode generation probability is heterogeneous. In the simpler case where the barcode generation probability ρ is homogeneous (as in the case of introducing lineage barcoding with lentivirus[20,21]), we have $M \leq 1 + 2\alpha/\rho$, or $1 + 2\alpha K$, where $K = 1/\rho$ is the total available barcodes in the barcoding library.

Given an allele bank, the maximum number of clones that it could reliably label at a fixed FDR α (Figure S3D) is given by $\max_{\rho_*} \min\left\{\frac{2\alpha}{\langle\rho|\rho \leq \rho_*\rangle}+1, N(\rho_*)\right\}$, where $N(\rho_*)$ is the total number of observed alleles in the bank that have allele probability $\rho \leq \rho_*$.

Allele bank construction

We use the term "allele bank" to refer to a collection of observed alleles and their associated frequencies. An allele bank was generated from bulk RNA-seq libraries prepared for each of the three target arrays from ~300,000 granulocytes from each of the three DARLIN mouse replicates (Figure S3B). These libraries were prepared shortly (3 days) after Dox treatment, so that the abundance of alleles within the data would reflect the frequency of their generation in cells, with little contribution from differential expansion among labeled clones. In total, this allele bank contained ~110, 62, and 37 thousand alleles from each of the CA, TA, and RA loci, respectively (Figure S3C). We further inferred the generation probability of each detected allele in this data (see below). This DARLIN allele bank is available within the MosaicLineage package described above. For Cas9/CARLIN, we used the alleles from the allele bank reported in the original study,[19] which was also generated by performing bulk RNA assays from three mouse replicates, each with ~300,000 granulocytes. This Cas9/CARLIN allele bank was only used in Figure 3N in this study to select rare alleles for both Cas9/CARLIN datasets involved.

For an allele i from source $s \in$ {CA, TA, RA, Cas9/CARLIN}, we inferred its generation probability ρ_i^s by normalizing its total UMI count X_i^s across the three mouse replicates. Here, $\rho_i^s = \gamma_s X_i^s / \left(\sum_j X_j^s \right)$, where γ_s is a source-specific pre-factor to correct bias arising from insufficient sampling. We fixed $\gamma_{CA} = 1$, and determined γ_s for other allele sources by fitting it to ensure that the same allele generation probability gave roughly similar barcode homoplasy probability (see homoplasy probability inference) among three mouse replicates for each allele source (Figure S3I). For *de novo* alleles, i.e., alleles not detected in our allele bank, we assign them a probability of zero. So far, we have inferred the allele/barcode generation probability for each of the target arrays separately. When alleles from more than one locus are detected in single cells, we concatenated them into a single (joint) lineage barcode, with a probability $\rho = \rho^{CA} \rho^{TA} \rho^{RA}$, where ρ^{CA}, ρ^{TA}, and ρ^{RA} are the barcode generation probability associated with alleles detected from each of the CA, TA, and RA loci, respectively. The inferred allele generation probabilities from our granulocyte-derived allele bank behave as expected across different tissues: alleles detected in multiple replicates of other tissues on average have higher generation probabilities in the allele bank than those detected in a single replicate (Figure S3E).

We have estimated that at FDR $\alpha = 0.01$, our current allele bank enables to reliably label ~10^4 clones using just one target array, ~10^8 clones using two target arrays, and ~10^{12} clones if all three arrays are used (Figure S3D). Furthermore, *de novo* alleles, if detected, can also be used to reliably label clones, which greatly expands the barcoding capacity of DARLIN. Indeed, ~80% of sampled barcodes from the TA locus in a new experiment will be *de novo* alleles (Figure S5D).

Finally, we note that the inferred allele generation probability ρ_i^s is almost certainly overestimated for the rarest alleles within the bank, due to our small sampling size in the allele bank. As a result, the number of identified alleles that can reliably label individual clones at a given FDR cutoff are likely more conservative (i.e., lower) than the true value. Indeed, as more DARLIN datasets are produced in systems with minimal clonal expansion (such as granulocytes), it will be possible to identify more DARLIN alleles that can be used for reliable clonal labeling in a given experiment.

Homoplasy probability inference

Here, we describe our method to infer the intrinsic homoplasy probability $\overline{P}_i^s$ of an allele i from source s by considering both the total UMI count X_i^s of this allele across the three replicates, and the number of mouse replicates N_i^s in which this allele was observed (Figure S3C). We define homoplasy probability of an allele to be the likelihood that this allele, already detected, will be observed in >1 mice of the three mouse replicates. Below, we will drop s in our notations just for simplicity.

Although X_i and N_i both positively correlate with the probability of an allele to be generated multiple times in a lineage-tracing experiment, thus resulting in barcode homoplasy, these two quantities may be influenced by target array expression or library sequencing depth. In fact, we observed that the array at the TA locus has ~3-fold greater expression compared to those at either the CA or RA loci. Our goal is to infer the intrinsic homoplasy probability $\overline{P}_i$ that accounts for the difference in both the target-array expression and the library sequencing depth between the CA, TA, and RA loci. We exclude unedited alleles in this analysis as their UMI counts depend on editing efficiency, which may skew the analysis.

We assign a homoplasy score $\sigma_i = 0$ for allele i if $N_i = 1$ and $\sigma_i = 1$ if $N_i > 1$. Then, to mitigate sampling noise, we average σ_i among alleles with the same UMI count X_i to obtain a smoothed score P_i, which we refer to as the observed homoplasy probability (for a large X, there may be few alleles with the same UMI count, and we resort to local averaging among alleles with a similar X). As expected, we find that P increases with X, and furthermore fully spans the interval between 0 and 1 for all three loci. However, the rate by which P increases with respect to X was different for all three loci (Figure S3F), presumably due to some combination of differences in expression and sequencing depth between these loci.

We account for sequencing depth by normalizing the UMI count of each allele with the total count of edited UMIs from a given locus. This is achieved by studying the relationship between P and the cumulative UMI fraction $\zeta(P) = \left(\sum_i X_i H(P_i < P) \right) \Big/ \sum_{i'} X_{i'}$, i.e., the fraction of UMIs associated with edited alleles that have an observed homoplasy probability $<P$, where $H(x) = \{1$ if the

condition x is true; 0 otherwise}. Indeed, we find that alleles from CA and RA loci, which have comparable expression (Figure S2H) yet different sequencing depths (Figure S3F), have the same relationship between P and ζ (Figure S3G), while alleles from TA have higher P at the same ζ due to their higher expression (Figure S2H).

We expect that, after accounting for expression differences to obtain the intrinsic homoplasy probability $\overline{P}$, alleles from the three loci should have the same relationship between $\overline{P}$ and ζ, i.e., $\overline{P}_{CA}(\zeta) = \overline{P}_{TA}(\zeta) = \overline{P}_{RA}(\zeta)$ (Figure S3H). For simplicity, we use the expression of alleles from the CA or RA loci as a reference. In this definition, the observed homoplasy probability in the CA or RA loci is the intrinsic homoplasy probability, i.e., $P_{CA}(\zeta) = \overline{P}_{CA}(\zeta)$ and $P_{RA}(\zeta) = \overline{P}_{RA}(\zeta)$. We introduce a mapping function $f(\cdot)$ that can map $P_{TA}(\zeta)$ to $\overline{P}_{TA}(\zeta)$, i.e. $f(P_{TA}(\zeta)) = \overline{P}_{TA}(\zeta)$. Combining the above relations, the mapping function can be solved from $f(P_{TA}(\zeta)) = P_{CA}(\zeta)$. We find that $f(x) \approx x/3$, which is consistent with the fact that alleles from the TA locus have about 3 times the expression of those from CA or RA loci (Figure S2H).

Clone identification: practice

Below, we discuss the choice of barcode generation probability cutoff ρ_* in each case.

1) For technical evaluation of single-cell lineage coverage (Figures 3N, 6C, and S6C), we wanted to obtain a metric that is independent of the sampling size. Therefore, we used the same cutoff $\rho_* = 2\times10^{-5}$ across different experiments and for alleles from different sources (CA, TA, RA or Cas9/CARLIN) to obtain the fraction of rare alleles. This cutoff was selected as we obtained ~10% trustable clonal cells in the published Cas9/CARLIN single-cell dataset (Figure 3N), consistent with the report in the original paper, suggesting a similar filtering stringency.
2) For Figures 4 and S4, we set $\rho_* = 2\times10^{-5}$ (same as above), which was derived using our above equation using $M = 1000$ clones and $\alpha = 0.005$. In this application, we also excluded alleles detected in more than one mouse replicate in our allele bank. For Figures 6, 7, S6, and S7 (excluding Figures 6C and S6C), we set $\rho_* = 2\times 10^{-4}$, which was derived with $M =$ 100 clones and $\alpha = 0.001$ (We detected ~100 clones per embryo when induction at E10). In both applications, we integrated alleles detected in single cells across CA, TA, and RA loci to determine which cells belonged to which clones. To do this, we constructed a graph in which each node represents a cell, with vertices connecting pairs of cells that share an allele (from CA, TA, or RA) with generation probability $\rho < \rho_*$. This procedure yielded a graph composed of many isolated subgraphs, with each subgraph corresponding to a different clone (where a clone is defined as a set of cells derived from a single progenitor). Finally, to further rule out the possibility that any of the retained clones were caused by the same allele being generated independently in multiple unrelated founder cells, we only considered clones having ≤ 3 different alleles from either CA, TA, or RA for further analysis. This cutoff is selected considering that a cell could have three different alleles if CA, TA, and RA alleles are all edited.
3) To quantify the circulation of hematopoietic cells (Figures 5 and S5), we applied the most stringent criteria, by i) using *de novo* alleles not present in *either* the three replicate mice in our allele bank, or six other mice profiled in the course of our study; and ii) the allele complexity was ≥ 4 (except in Figure 5E). Allele complexity was calculated as the sum of both the total number of mutation events and the total number of inserted base pairs within an allele, such that an allele with only a deletion would have an allele complexity of 1, and an allele with both a deletion and a 2-bp insertion would have an allele complexity of 4 (2 events + 2-bp insertion).

Clonal analysis

Unless otherwise stated, we excluded clones with only one cell (from single-cell lineage tracing) or one UMI (from bulk lineage tracing) in downstream analysis in Figures 4, 5, 6, and 7, because such clones cannot be used to assess lineage relationships between cell types or bones.

Clonal fate bias test

To evaluate the statistical significance of the fate bias of a clone (Figure 4C), we compared the observed fate partitioning of a clone with that expected by chance, given the distribution of fates of all observed cells. Specifically, for a clone with in total M cells in the differentiated cell states, among which m of them belong to a fate cluster of interest, we performed a one-sided hypergeometric test from the values (M, m, N, n) to obtain the p-value of a clone, where N is the total number of cells in the differentiated cell states across all clones, and n is the corresponding cell number within the targeted fate cluster.

Clonal coupling analysis

We computed a normalized correlation coefficient, which quantifies how often two cell types jointly appear within the same clone. We have applied this approach to both single-cell (Figure 4D), and bulk (Figures 5C and 5G–5I) lineage tracing data. When applied to bulk lineage tracing data, each UMI was inferred to correspond to a distinct cell, and so the term “cell” is used in the following to refer either to a UMI or a cell-level barcode depending on the experiment. We represent the cell-by-clone matrix as $\boldsymbol{X}$, with the entry $X_{ij} \in \{0, 1\}$, where a value of 1 indicates that cell i belongs to clone j. This matrix is used to aggregate all cells within each population (e.g., cell type) and the same clone to obtain matrix A, where each A_{kj} represents the number of cells from cell type k belonging to clone j. A_{kj} constructed from bulk or single-cell lineage tracing data from the same biological system can be related to each other after adjusting for library depth differences. Since different cell types may be sampled at different depths, we first row-normalize the matrix A_{kj} by the total number of cells of each cell type k to account for sampling heterogeneity between cell types (so that each row adds to

1). Besides, because each clone represents an independent measurement and therefore should contribute equally to the coupling calculation, we then normalize the matrix per column to calculate within each clone the fraction of cells in each cell type (clone normalization). This clone normalization was performed first among HSPCs and then among downstream fates, so that the probability mass of a clone is conserved over different differentiation stages. The above two normalization procedures yield the normalized clonal matrix $\overline{A}_{kj}$. The transpose of the $\overline{A}_{kj}$ matrix is shown as heatmaps in Figures 4B, S4B, and S5B. To calculate the clonal coupling from this matrix, we follow the approach in CoSpar.[49] Briefly, the clonal coupling is calculated as $Y_{kk'} = \sum_j \overline{A}_{kj}\overline{A}_{k'j}$, and then subsequently normalized according to $\overline{Y}_{kk'} = Y_{kk'}/\sqrt{Y_{kk}Y_{k'k'}}$ to obtain the reported clonal coupling scores between cell types k and k' (Figure 4D).

To evaluate the statistical significance of an observed clonal coupling computed from above, we compared it with that from 10,000 randomized clonal matrices to obtain the one-sided p-value. The randomized clonal matrices were generated to preserve the same cell number in each cluster, and the same clone size distribution among HSPCs as well as downstream fates. To achieve this, we first filtered the cell-by-clone matrix X to include only cells having valid clonal IDs, then split the matrix into two sub-matrices: $X(t_0)$, which contains only cell states from HSPCs, and $X(t_1)$, which contains all remaining cell states. We randomly shuffled $X(t_0)$ and $X(t_1)$ along the row axis, and then concatenated these two matrices to obtain the final randomized clonal matrix X. We updated the CoSpar package during this study to implement the above clonal analysis: the clonal coupling analysis can be achieved with the *cospar.tl.fate_coupling* function, with the p-value generated by *cospar.tl.pvalue_for_fate_coupling*.

Progenitor fate bias prediction

To infer the fate bias of HSPCs (Figure 4F), we manually defined HSPCs as the early population (t_0) and the rest cells as the later population (t_1) in order to apply CoSpar to infer the transition probability matrix from HSPCs to any cell from the non-HSPC population. We used *cospar.tmap.infer_Tmap_from_multitime_clones* to infer the transition matrix. We then ran *cospar.tl.fate_map* with default parameters to infer the progenitor probability of each HSPC to major cell types (MkP, Ery, Mon, or Neu) from the computed transition matrix. The progenitor probability, normalized by the maximum value among progenitors, reports the probability of a given cell type originating from a given HSPC.

Shared clone fraction

We define $\varphi(x, y)$ as the fraction of clones that are detected in a given cell type x from one bone y and shared in the same cell type from other bones. We applied it to the bulk lineage data from Figure 5B, where we know for each UMI the associated lineage barcode, the cell type (through FACS), and bone location (because sequencing amplicons from each dissected region were prepared with a different library index). A clone was therefore identified as a collection of UMIs sharing a single lineage barcode, which might include UMIs that each correspond to different cell types and/or different bones. Denote $n_{shared}(x, y)$ as the number of clones shared between cells of cell type x from bone y and those from any other bones, and $n_{total}(x, y)$ as the total number of clones with cells of cell type x and bone y. Then, shared clone fraction $\varphi(x,y) = n_{shared}/n_{total}$.

Single-cell transcriptomic analysis

We generated count matrices from the 10X scRNA-seq data using Cell Ranger v.7.0.0. For transcriptomic data from Camellia-seq, we adapted a bioinformatic pipeline developed earlier to analyze STRT-seq data.[59] We used the mouse genome mm10 as the reference genome.

For quality control, we filtered the count matrices to remove cells either with <700 detected genes or with <1,500 total UMI counts. We next removed genes detected in <2 of the remaining cells. Finally, we removed cells for which mitochondrial genes contributed >10% of total UMIs. The count matrices were then normalized such that the total UMI counts corresponding to each cell were 10,000.

For dimensionality reduction (Figures 3I, 7C, and S6H), we adapted the pipeline described in CoSpar.[49] Briefly, we selected highly variable genes, regressed out the effects of cell cycle (unless otherwise stated), and used the top 40 principal components to build a k-nearest neighbor (k-NN) graph, with *n_neighbors* =20. Finally, we performed UMAP to generate the two-dimensional embedding.

Chromatin-accessibility and DNA-methylation analysis

Reads from scBS-seq libraries were aligned to the bisulfite-converted mm10 mouse genome using Bismark v.0.23.0 in a paired-end, non-directional mode. Then the unmapped reads were aligned again in a single-end, non-directional mode. We used the *coverage2-cytosine* script from Bismark with the –nom-seq option to generate CpG report files for ACG and TCG trinucleotides and GpC report for GCA, GCC, and GCT trinucleotides. In each cell, the DNA methylation fraction (CpG fraction) at a given CpG site was determined by the ratio of the number of methylated CpG reads to the number of unmethylated CpG reads. The accessibility fraction or GpC fraction was computed similarly. For quality control, we discarded cells with either <200,000 unique CpG sites, or <2,000,000 unique GpC sites.

We then projected the raw CpG or GpC fraction data onto a set of informative genomic regions for downstream analysis. For chromatin accessibility (GpC), we used the accessible regions (narrowpeak output from MACS2) from published bulk ATAC-seq data of HSCs[54] as the genomic feature sets, while for DNA methylation (CpG), we used the low-methylation regions identified in the pseudobulk DNA methylation data of bone marrow HSCs (CpG fraction cutoff set at the lowest 15 percentile). We averaged the CpG (or GpC) fraction over each of these pre-selected genomic regions, which we call features, to obtain the cell-by-feature rate matrix.

To perform dimensionality reduction with either the DNA-methylation or chromatin-accessibility cell-by-feature matrices, we first calculated the cell–cell similarity matrix S, with each element S_{ij} given by the Pearson correlation between cells i and j over all features. We then selected the top 10 eigenvectors of S to generate a k-NN graph with *n_neighbor* = 10. Finally, we performed UMAP to generate the two-dimensional embedding.

Clonal memory analysis

Intra-clone similarity score

To calculate intra-clone similarity scores with respect to each of the gene-expression, chromatin-accessibility, and DNA-methylation data (Figures 7F, 7G, and 7H), we first computed a cell-cell similarity matrix S using all features within a given data type. The intra-clone similarity score for each clone was then calculated as the average over all $(n^2 - n)/2$ pairwise similarity scores among the n cells associated with each clone. We also generated randomized clones by shuffling the cell-by-clone matrix X_{ij} along each column (clone) to preserve the original clone size distribution and computed the intra-clone similarity for these randomized clones, repeating this procedure 1000 times. Similarity scores from both observed and randomized clones were rescaled together within each data modality from a given mouse sample to span the interval [0,1]. We computed a p-value using Wilcoxon rank-sum test by comparing the observed intra-clone similarity scores of all clones from a data modality with those from corresponding randomized clones.

Significance of clonal memory per clone

To obtain the p-value of a clone regarding its clonal memory based on a certain data modality (Figure S6A), we retained the raw similarity scores between any two pairs of cells within a clone, and compared them with those between cells that do not share a clonal relationship. We then performed the Wilcoxon rank-sum test to derive a p-value, followed by Benjamini–Hochberg correction.

Memory scores

Memory scores O in Figure 7J were calculated as the difference between the average intra-clone similarity scores of the observed ($\underline{obs.}$) and randomized ($\underline{rand.}$) clones, normalized by the standard deviation of the intra-clone similarity scores of the randomized clones (std. rand.). That is $O = (\underline{obs.} - \underline{rand.})/(\text{std. rand})$.

Supplemental figures

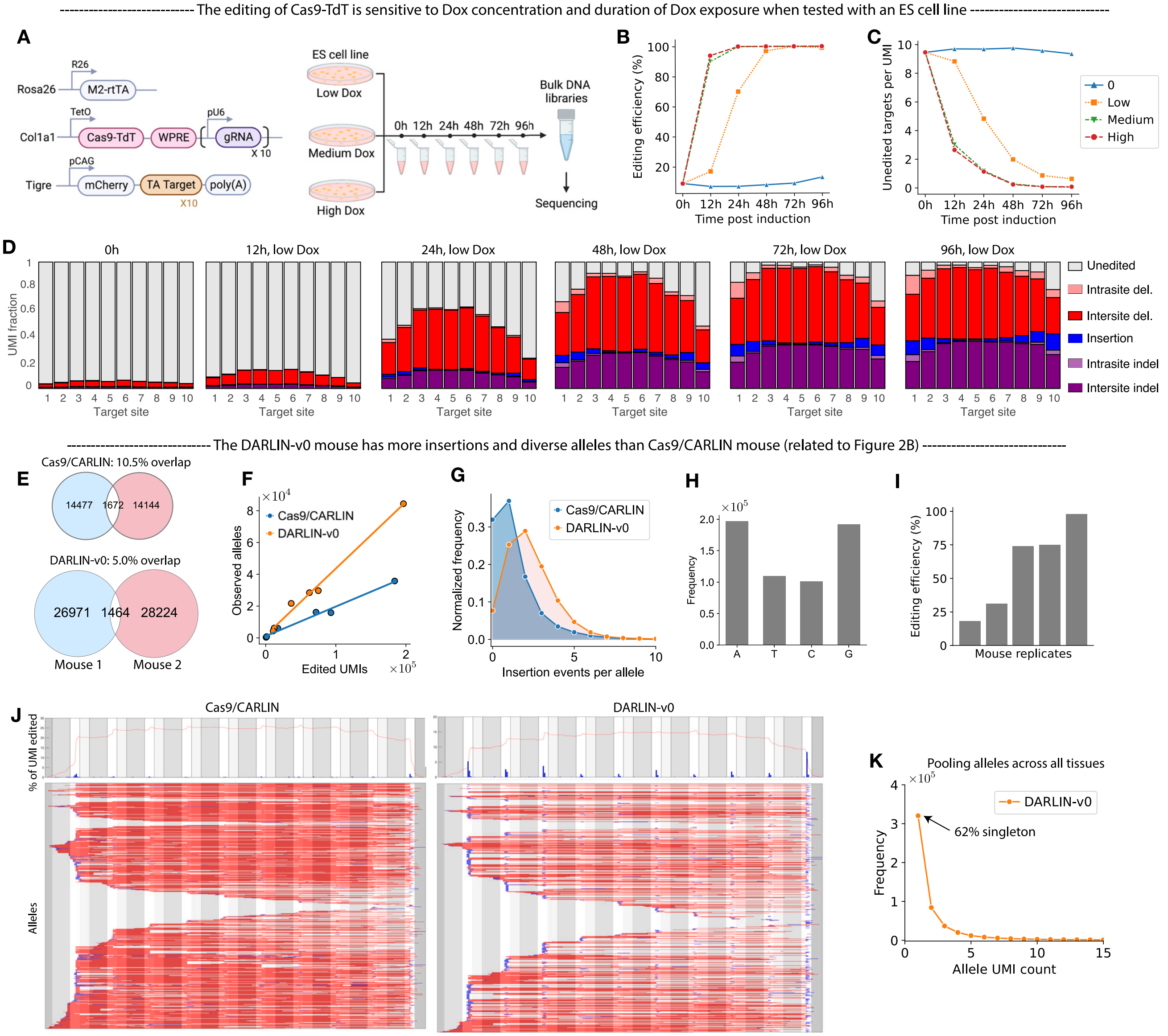


Figure S1. Characterizing the TetO-Cas9-TdTmESC line and the DARLIN-v0 mouse line, related to Figure 2

(A–D) Characterizing the TetO-Cas9-TdT mESC line.

(A) Genetic scheme of the mESC line (left) and the experimental scheme to test the sensitivity of its target-array editing to Dox induction (right). Note that the target array was integrated at the *Tigre* locus in this mESC. Cells were exposed to 0, low (0.04 µg/mL), medium (0.2 µg/mL), and high (1 µg/mL) Dox concentrations for 0, 12, 24, 48, 72, and 96 h, respectively.

(B and C) Relationship between Dox incubation time and the fraction of recovered target arrays (UMIs) that were edited (B) or the number of unedited target sites per recovered target array (C) at each of the four Dox concentrations.

(D) Relative UMI fraction of editing patterns across the ten target sites, for each length of time of Dox incubation at 0.04 µg/mL.

(E–G) Comparison of the alleles generated in the DARLIN-v0 mouse line with those in the Cas9/CARLIN mouse line, using the bulk RNA data from Figure 2B. Bulk RNA data from granulocytes were used to generate all the following figures except (I).

(E) Venn diagrams of the allele overlap between mouse replicate from the Cas9/CARLIN (top) and DARLIN (bottom) mouse lines.

(F) The relationship between the number of observed alleles and the inferred number of edited cells (i.e., UMIs).

(G) Distribution of the number of insertion events per observed allele.

(H) Frequency of all four DNA nucleotides among all inserted sequences identified in the edited alleles observed with the DARLIN-v0 mouse line.

(I) Fraction of cells (UMIs) from granulocytes that were edited across DARLIN-v0 mouse replicates.

(legend continued on next page)

(J) Maps of editing positions observed for either the Cas9/CARLIN (left) or DARLIN-v0 (right) mouse line. The upper panel shows the frequency of insertions (blue bars) or deletions (red lines) at each position of the unedited target-array sequence. For the insertions, the position corresponds to position immediately 5′ of the inserted sequence. The bottom panel depicts individual observed alleles, also mapped to the unedited target array. Deletions are represented by red lines spanning each deleted region. Insertions are represented by blue lines, each of which begins at the position immediately 5′ of the corresponding inserted sequence, with a length representing the corresponding number of nucleotides inserted.

(K) Histogram of allele UMI counts when aggregating alleles from all bulk RNA lineage-tracing libraries generated from DARLIN-v0 mice in Figure 2B. This aggregated dataset corresponds to the rightmost point of Figure 2M.

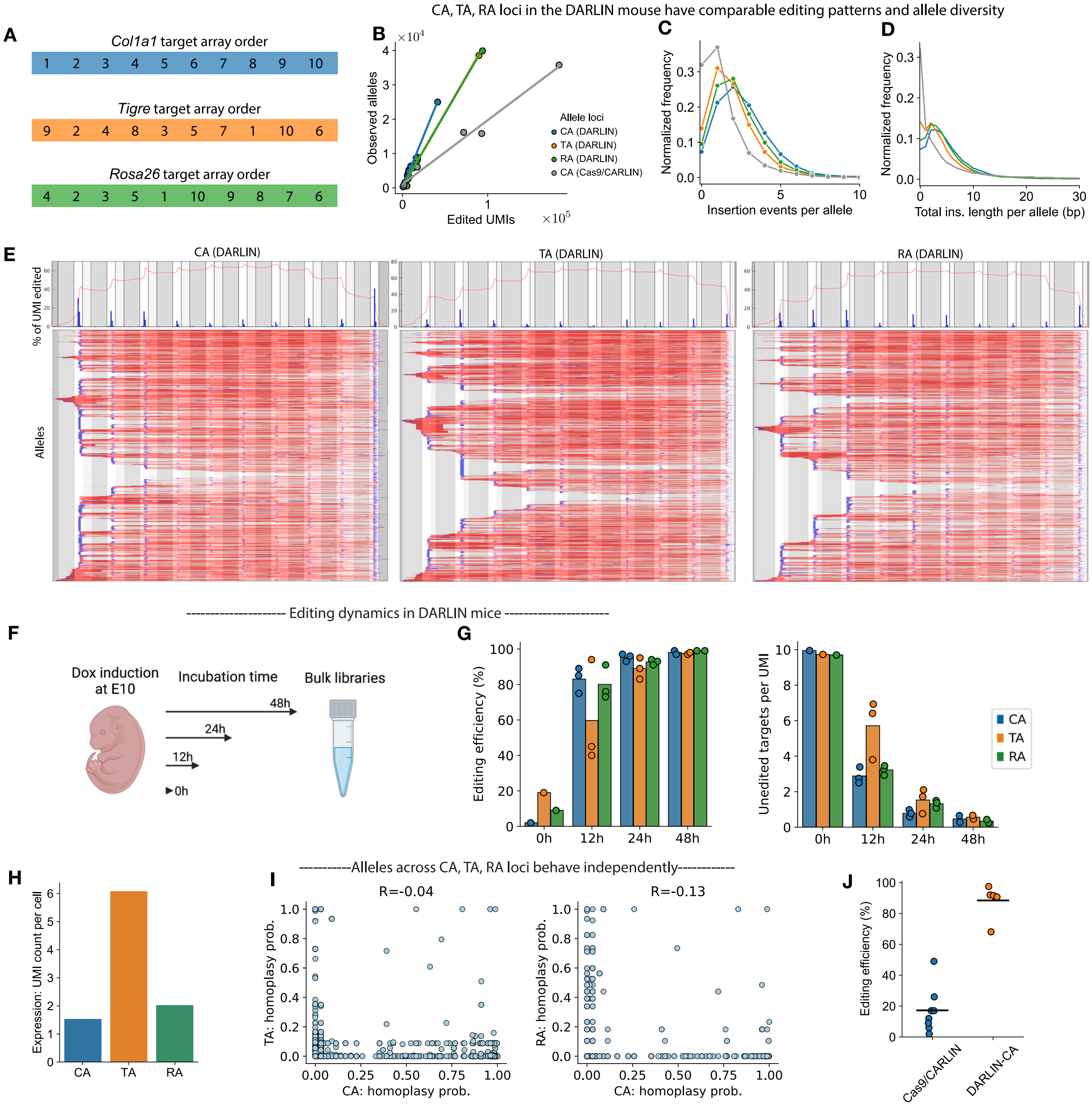


Figure S2. Additional characterization of the DARLIN mouse line, related to Figure 3

(A) Target-array order in the *Col1a1* (left), *Tigre* (middle), and *Rosa26* (right) loci, respectively.

(B–E) Comparison of alleles from the CA, TA, or RA loci (related to Figure 3B): the relationship between the number of observed alleles and the number of edited cells (UMIs) (B), distribution of either the number of insertion events (C) or the total insertion length (D) per allele, and maps of individual alleles (E).

(F) Experimental scheme to determine the time needed for target-array editing in DARLIN. Mouse embryos at E10 were treated with Dox for 0, 12, 24, and 48 h, respectively. All cells from these embryos were immediately profiled with bulk sequencing for editing in the CA, TA, and RA loci after treatment.

(G) Relationship between the fraction of cells (UMIs) that were edited (left) or the unedited targets per UMI (right) and the duration of Dox treatment.

(H) Average expression of the transcribed target array from either the CA, TA, or RA loci. The data are from the targeted amplification of each target array in the experiment associated with Figures 3H–3N.

(I) Scatter plot showing homoplasy probability of two alleles jointly detected within the same cell: CA and TA allele (left); CA and RA allele (right). For each of the two alleles found in the same cell in the experiment associated with Figure 3I, we queried the allele bank to obtain the corresponding pre-inferred intrinsic homoplasy probability $\overline{P}$. See Figure S3 and STAR Methods for details about the allele bank and homoplasy probability.

(J) Editing efficiency of the target array from fetal liver cells in the Cas9/CARLIN or DARLIN mouse line. Each point represents a mouse embryo replicate. Only CA alleles were amplified from the DARLIN mouse line.

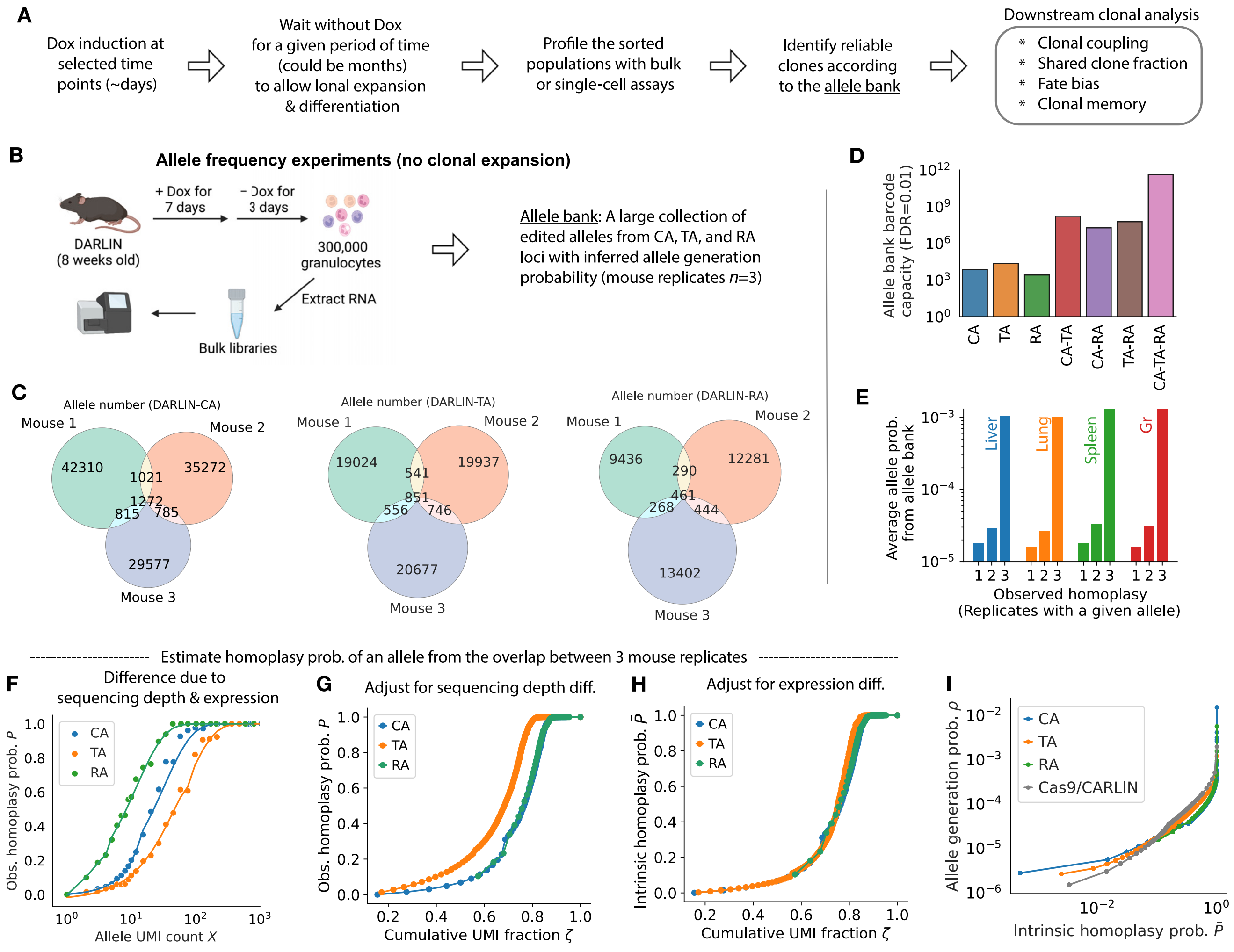


Figure S3. Allele bank construction and allele probability inference, related to Figure 3

(A) Flowchart for a lineage tracing experiment with the DARLIN mouse line and its data analysis.

(B) Experimental scheme to generate a large allele dataset in DARLIN mice that measured the allele generation probability. This dataset was used to construct the allele bank.

(C) Venn diagram of the allele overlap between the three mouse replicates generated in (B), using alleles from the CA, TA, and RA loci.

(D) Maximum number of clones that the allele bank could reliably label at a false discovery rate (FDR) of 0.01, when using different combinations of the CA, TA, and RA loci.

(E) Relationship between the observed homoplasy of an allele detected in a given tissue or cell type and the average predicted allele generation probability from the allele bank (constructed using granulocytes). The observed homoplasy of an allele refers to the number of replicates (1, 2, or 3) for which that allele was detected in a given tissue or cell type. *De novo* alleles (alleles not found in the allele bank) were excluded in this analysis. Data are from the experiment described in Figure 2B.

(F–H) Estimation of the homoplasy probability of an allele from the observed overlap between three mouse replicates. See STAR Methods section on homoplasy probability inference for details.

(F) Relationship between the observed homoplasy probability of an allele and its corresponding UMI count.

(G) Relationship between the observed homoplasy probability P and the cumulative UMI fraction, i.e., the fraction of UMIs associated with edited alleles that have an observed homoplasy probability $< P$.

(H) Relationship between the intrinsic homoplasy probability $\overline{P}$ and the cumulative UMI fraction.

(I) Relationship between the inferred generation probability of an allele and its intrinsic homoplasy probability.

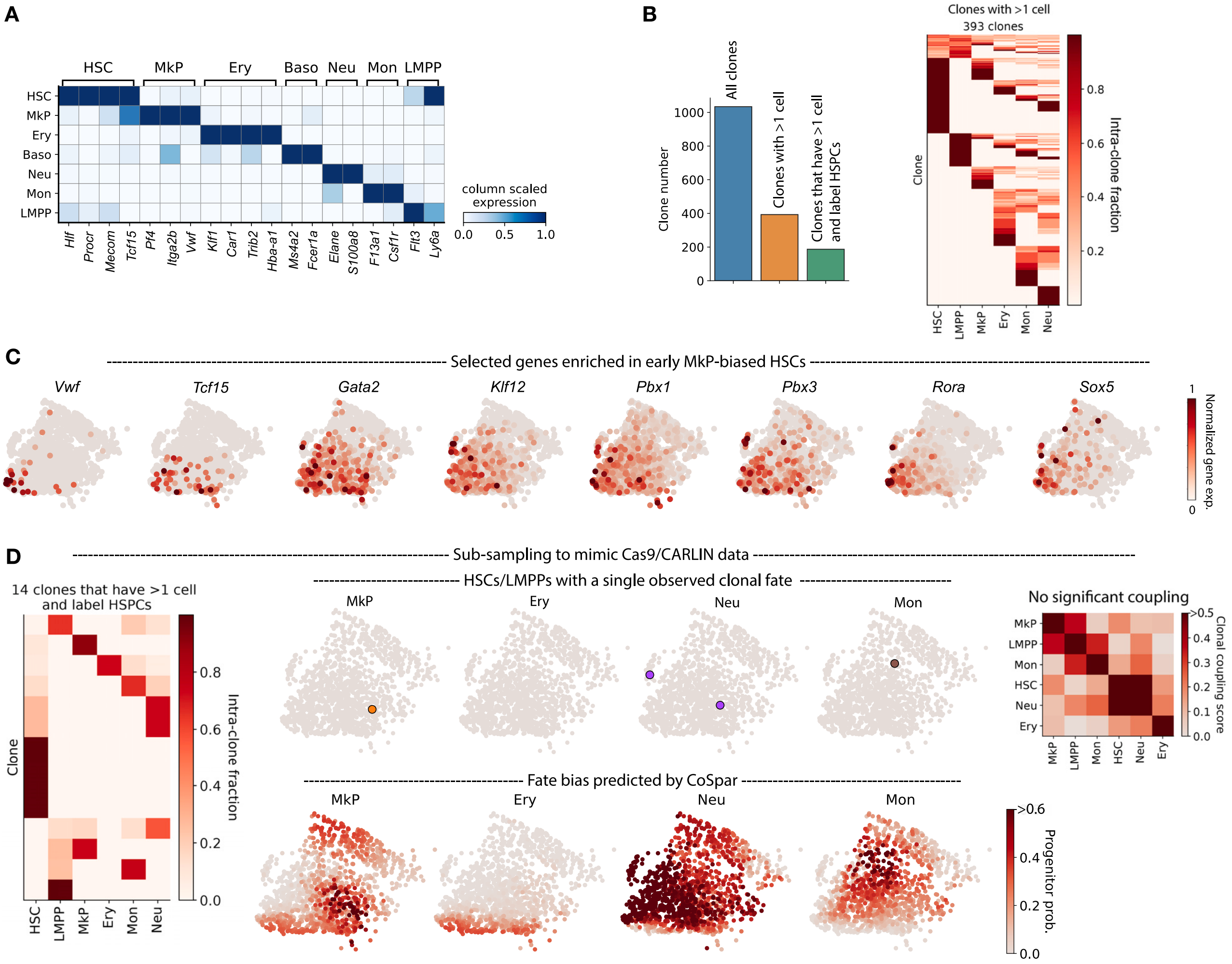


Figure S4. Additional quantification of single-cell lineage-tracing data from skull bone marrow, related to Figure 4

(A) Heatmap showing the expression of cell-type-specific marker genes (columns) in each annotated cell type (rows) using data from Figure 3I. Expression values were column-wise normalized by the highest value in each column.

(B) Characterization of the clones detected in skull-derived bone-marrow data with respect to the number of clones satisfying increasingly stringent criteria (left) and the cell-type composition of the 393 clones with more than one cell (right).

(C) Expression of MkP bias-associated genes in HSPCs, in addition to Figure 4I. These selected genes either play an important role in megakaryocytes (e.g., *Vwf*) or are transcription factors with a putative role in MkP bias.

(D) Clonal fate-bias analysis (same as Figures 4B–4F) with cells from the skull-derived bone marrow after sampling the lineage-tracing data to resemble that of the Cas9/CARLIN mouse line. We used only rare alleles from the CA locus (comprising 27% of the total rare alleles) to generate clones and further down-sampled these clones to 44% × 69% = 30%, where the 44% accounts for the reduced editing efficiency in the Cas9/CARLIN mouse (~35%) in comparison to that of the DARLIN mouse (~80%), giving the relative fraction 35%/80% = 44%, whereas the 69% is due to the reduced rare-allele fraction among edited alleles in the Cas9/CARLIN mouse (~55%) than in DARLIN (~80% when considering only alleles only from CA locus), leading to the relative fraction 55% / 80% = 69%. The sampling parameters were derived from Figure 3N. None of the clonal coupling scores (upper right) were statistically significant (threshold: p = 0.05).

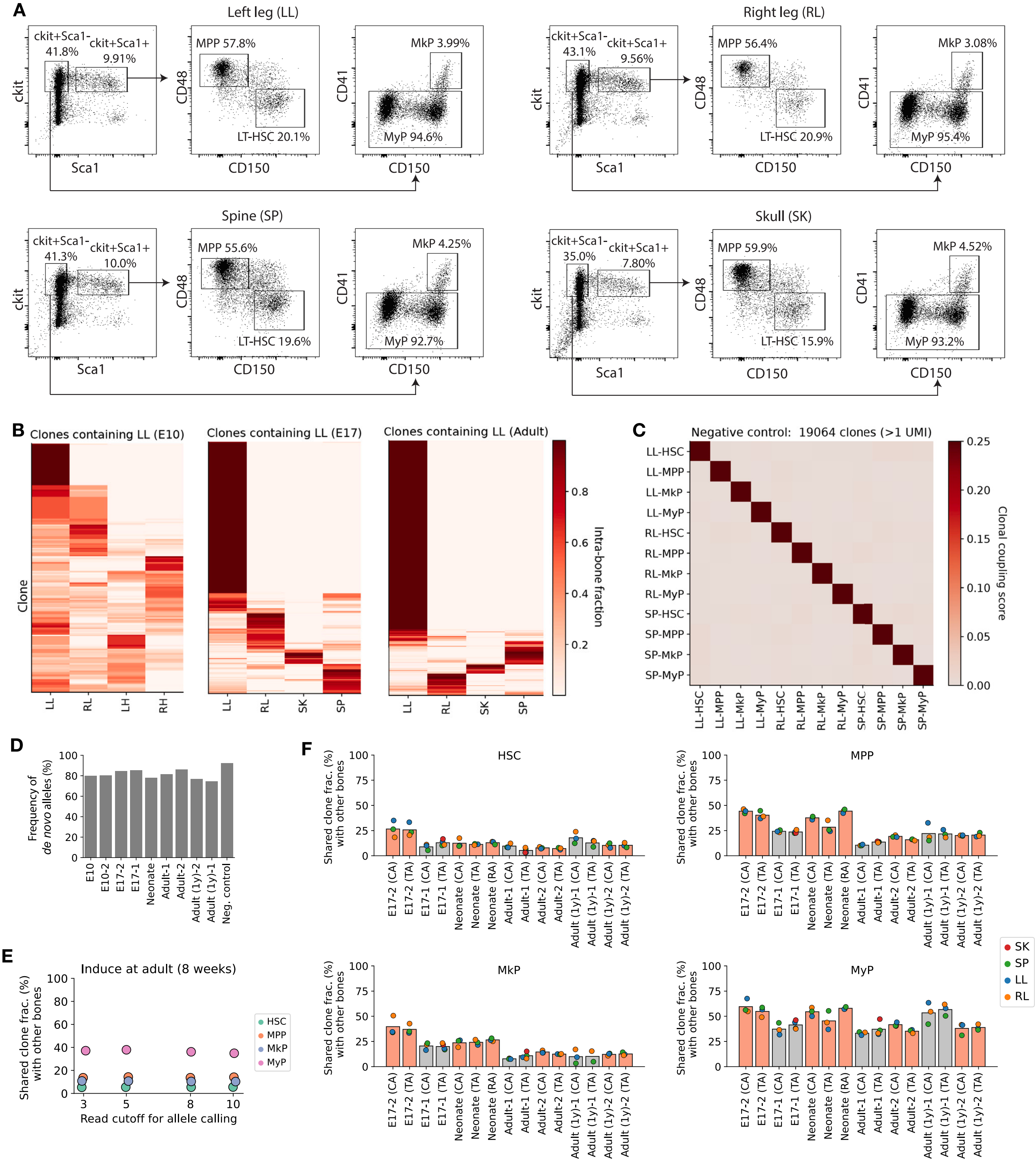


Figure S5. Additional analysis of inter-bone lineage relationships, related to Figure 5

(A) FACS sorting strategies for obtaining LT-HSC, MPP, MkP, and MyP.

(B) Clonal profiles of hematopoietic cells from DARLIN mice induced at E10.0 (left), at E17.0 (middle), and in adulthood (right). Otherwise, as in Figure 4B. Only clones labeling the left leg bone (LL) were shown.

(C) Heatmap of the clonal coupling scores between different cell types and bones, using data from a negative control mouse, with which lineage barcodes were induced in the adult stage and profiled 3 days later, sorting for bone-marrow-derived LT-HSC, MPP, MkP, and MyP cells from the LLs, right leg bones, and spine.

(legend continued on next page)

(D) Percentage of *de novo* alleles in the lineage-tracing data corresponding to each of the collected mouse samples. A *de novo* allele is an allele that was not found in our large allele bank.

(E) The relationship between the observed shared clone fraction and the reads-per-UMI threshold for calling alleles, plotted for each of the four cell types considered in Figure 5D.

(F) Shared clone fraction with respect to HSC, MPP, MkP, and MyP inferred using alleles profiled from each of the CA, TA, or RA loci within the same mouse, across different induction stages. The TA data are reproduced from Figures 5J–5M to compare the technical consistency between different loci within the same mouse. Data from the same mouse are plotted with the same color. Otherwise, same as Figure 5J.

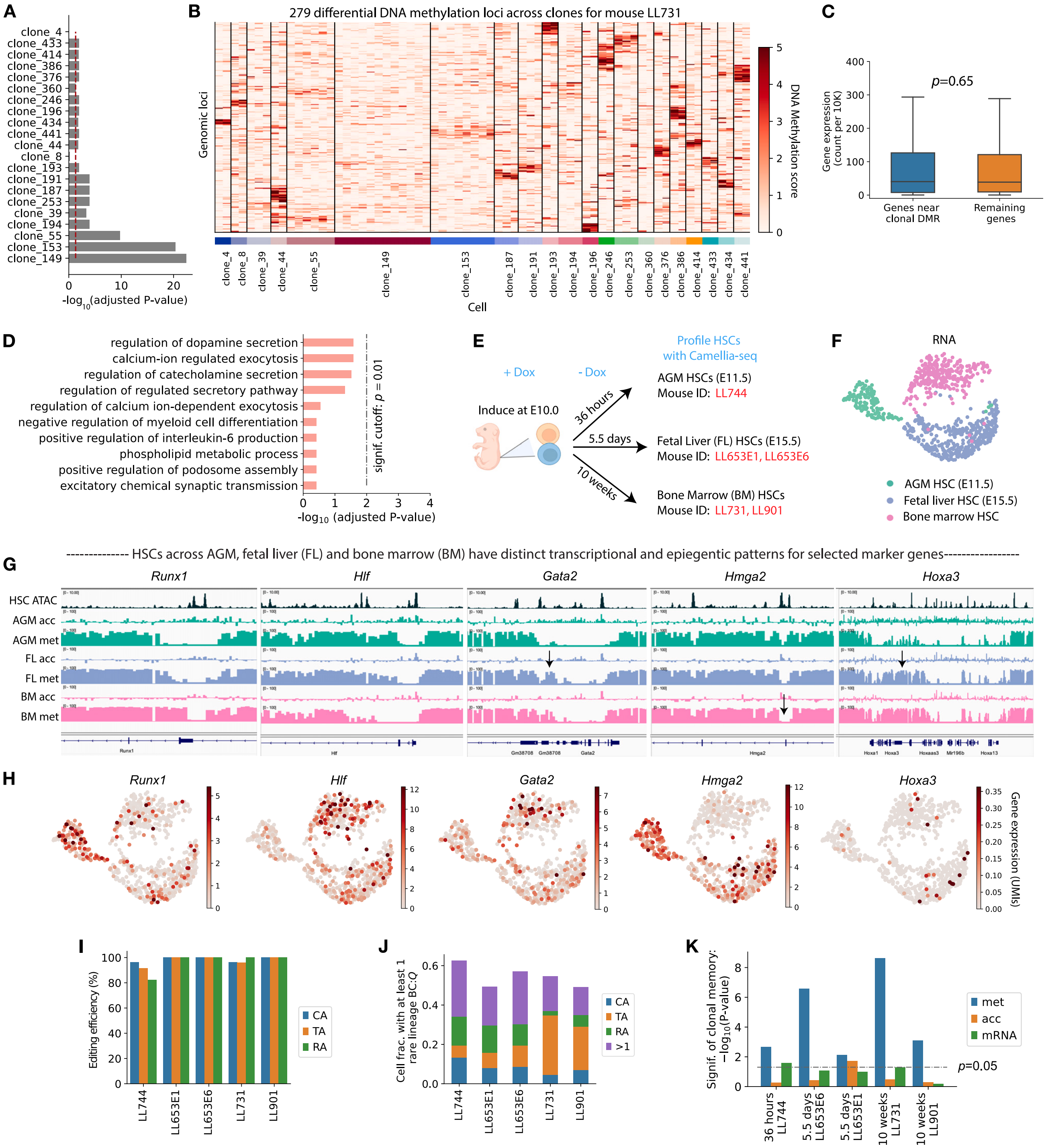


Figure S6. Characterizing Camellia-seq HSC datasets across stages, related to Figures 6 and 7

(A) Adjusted p values of intra-clone similarity with respect to DNA methylation for all the 21 observed clones with $\geq$ 2 cells. The red dashed line indicates p = 0.05. Significant intra-clone similarity was observed for 19 out of all 21 clones.

(B) Differential DNA-methylation regions (DMRs) between clones with >1 cells in mouse LL731. Each column corresponds to an individual cell, and each row corresponds to one of 279 genomic loci that were identified as significantly varying ($p \leq 0.05$, one-way ANOVA, with Benjamini-Hochberg correction). The DNA-methylation score is obtained by smoothing the CpG fraction on the full PCA space as described in Kremer et al.[70]

(C) Box plot of expression associated with genes closest to the 276 clonal DMRs identified in (B) (blue) and all remaining genes (yellow). The difference is insignificant between these two groups (p = 0.65, t test).

(D) Top 10 enriched GO terms associated with genes closest to 276 clonal DMRs. No statistically significant GO terms were found at an FDR cutoff of 0.01.

(legend continued on next page)

(E) Experimental scheme by which the Camellia-seq experiments were performed with DARLIN mice at different stages post induction at E10.0. LL731 is highlighted in Figures 6 and 7.

(F) UMAP embeddings of all post-filtering cells in the gene expression space, without cell-cycle correction. Cells are colored by their developmental stages.

(G and H) Examples of stage-specific chromatin accessibility, DNA methylation, and gene expression.

(G) Pseudobulk chromatin-accessibility and DNA-methylation data across HSC stages for selected genes: *Runx1*, *Hlf*, *Gata2*, *Hmga2*, and *Hoxa3*. The ATAC peaks of HSCs from Li et al.[54] are also shown. The arrows indicate genomic regions with differential regulation across developmental stages.

(H) Gene expression on the transcriptome UMAP embedding, related to (G).

(I) Fraction of cells with an edited target array (from CA, TA, or RA loci) for each of the collected mouse samples.

(J) Fraction of cells for which a rare allele was detected from at least one target array for each mouse sample, as measured by Camellia-seq. Otherwise, as in Figure 3L.

(K) Statistical significance (i.e., p value) of transcriptomic and epigenomic memory for each of the collected mouse datasets. For a given mouse sample and data modality, its corresponding p value was calculated as in Figures 7F–7H.

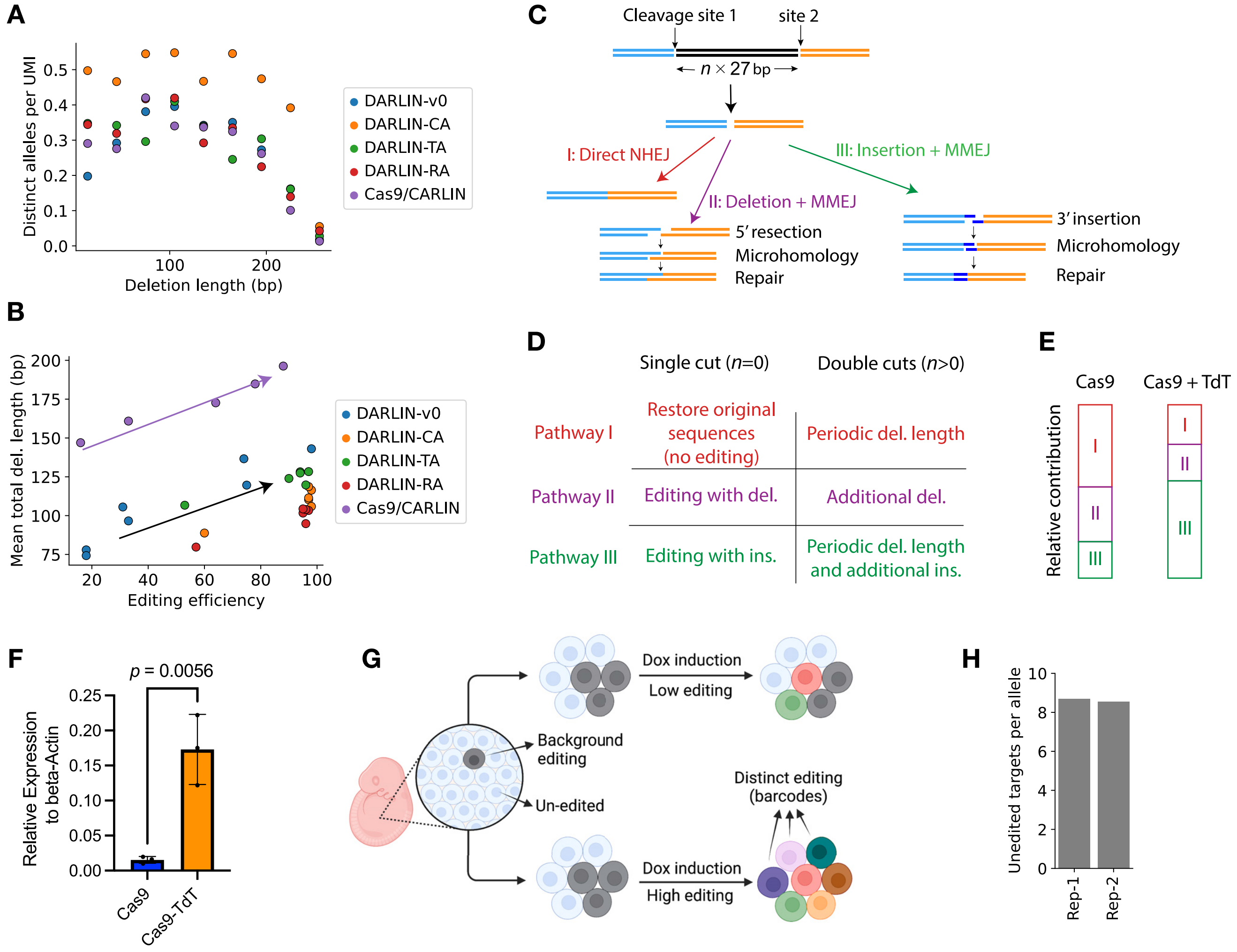


Figure S7. Additional characterization of DARLIN, related to Figures 2 and 3

(A) The relative abundance of unique deletion-only alleles as a function of deletion length. The relative abundances were calculated by dividing the number of unique alleles by total UMI count in each sample. Each dot represents a bulk allele library from 300,000 granulocytes extracted from a mouse replicate, where the Cas9/CARLIN and DARLIN-v0 datasets were generated from Figure 2B and DARLIN-CA, DARLIN-TA, and DARLIN-RA datasets from Figure 3B. To mitigate the influence of insertions on the quantification, we considered alleles with only deletions. The yield of distinct alleles decreases for large-scale deletions (i.e., >180 bp), which remove the majority of the targeting array.

(B) Relationship between the mean total deletion length among all alleles in a bulk sample and the corresponding editing efficiency of this sample. Otherwise, same as in (A). We observed shorter deletion lengths in DARLIN than Cas9/CARLIN at the same editing efficiency.

(C–E) Proposed model for the observed editing differences between Cas9 and Cas9-TdT.

(C) Illustration of three repair pathways (I, II, and III) following double-strand breaks (DSBs) generated by Cas9 or Cas9-TdT. A DSB can occur at a single target site (single cut) or two sites (double cuts), with a length of n×27 bp, where n is the number of deleted target sites in between. DSBs can be repaired through direct non-homologous end joining (NHEJ, pathway I), with no further insertions or deletions. These dual-site DSBs may also undergo further 5′ resection, before being resolved using exposed microhomology (2–3 bp) between the exposed 3′ sequences through microhomology-mediated end joining (MMEJ) (pathway II),[71] resulting in additional deletions. Finally, DSBs may undergo 3′ extension at either or both ends of the cleavage site. This could generate microhomology between these 3′-extended sequences and lead to DSB repair through MMEJ (pathway III), resulting in additional insertions.

(D) Outcomes when repairing a DSB associated with a single cut or double cuts using each of the three repair pathways. In the case of a single cut, pathway I restores the original sequence and leads to unsuccessful editing, pathway II generates edited alleles with additional deletions, and pathway III generates edited alleles with additional insertions. In the case of double cuts, pathway I leads to a periodic deletion length of n×27 bp, pathway II gives the rightward skew of deletion density observed at each periodic peak of deletion length, and pathway III generates alleles with periodic deletion length and additional insertions.

(E) Relative contribution of each repair pathway associated with either Cas9 (Cas9/CARLIN) or Cas9-TdT (DARLIN). Without TdT, pathway I is thought to be the dominant repair pathway.[72] We propose that TdT significantly increases the contribution from pathway III, which would lead to more insertions, fewer deletions (due to less contribution from pathway II), and increased editing efficiency (due to less contribution from pathway I that restores single-cut events into unedited alleles). The collective contribution from all three pathways in either DARLIN or Cas9/CARLIN is shown in Figure 2F, which is consistent with our model prediction.

(F) Gene expression of Cas9 or Cas9-TdT in all cells from Cas9/CARLIN and DARLIN embryos, respectively, quantified by qPCR following 24-h Dox treatment. The p value from t test is shown.

(legend continued on next page)

(G) Illustration of the impact of background editing in the scenario of low or high efficiency of Dox-induced editing. With high efficiency of Dox-induced editing, alleles arising from background editing may be further edited in a stochastic manner, such that each of the cells (colored in gray) carrying this background allele will carry a new and distinct allele following Dox induction, thereby mitigating the impact of background editing.

(H) Average number of unedited targets for alleles from background editing. Data were from granulocytes of two mouse replicates without Dox induction, generated in Figure 2B. It supports the possibility of further editing these background-generated alleles upon Dox induction.