Polylox barcoding reveals haematopoietic stem cell fates realized in vivo

Weike Pei1, Thorsten B. Feyerabend1, Jens Rössler2, Xi Wang2,3, Daniel Postrach1, Katrin Busch1, Immanuel Rode1, Kay Klapproth1, Nikolaus Dietlein1, Claudia Quedenau1, Wei Chen1, Sascha Sauer3, Stephan Wolf4, Thomas Höfer2,5 & Hans-Reimer Rodewald1

Developmental deconvolution of complex organs and tissues at the level of individual cells remains challenging. Non-invasive genetic fate mapping1 has been widely used, but the low number of distinct fluorescent marker proteins limits its resolution. Much higher numbers of cell markers have been generated using viral integration sites2, viral barcodes3, and strategies based on transposons4 and CRISPR–Cas9 genome editing5; however, temporal and tissue-specific induction of barcodes in situ has not been achieved. Here we report the development of an artificial DNA recombination locus (termed Polylox) that enables broadly applicable endogenous barcoding based on the Cre–loxP recombination system6. Polylox recombination in situ reaches a practical diversity of several hundred thousand barcodes, allowing tagging of single cells. We have used this experimental system, combined with fate mapping, to assess haematopoietic stem cell (HSC) fates in vivo. Classical models of haematopoietic lineage specification assume a tree with few major branches. More recently, driven in part by the development of more efficient single-cell assays and improved transplantation efficiencies, different models have been proposed, in which unilineage priming may occur in mice and humans at the level of HSCs8–10. We have introduced barcodes into HSC progenitors in embryonic mice, and found that the adult HSC compartment is a mosaic of embryo-derived HSC clones, some of which are unexpectedly large. Most HSC clones gave rise to multilineage or oligolineage fates, arguing derived HSC clones, some of which are unexpectedly large. Most HSC clones gave rise to multilineage or oligolineage fates, arguing against unilineage priming, and suggesting coherent usage of the potential of cells in a clone. The spreading of barcodes, both after induction in embryos and in adult mice, revealed a basic split between common myeloid–erythroid development and common lymphocyte development, supporting the long-held but contested view of a tree-like haematopoietic structure.

To take advantage of available mutants expressing inducible Cre recombinase in specific cell types or developmental stages, we designed a DNA cassette, Polylox, for endogenous barcoding. Polylox is composed of ten loxP sites in alternating orientations spaced 178 base pairs (bp) apart11, thus allowing excisions and inversions (Fig. 1a). The intervening nine DNA blocks have unique sequences based on the AT2G21770 gene from Arabidopsis thaliana, serving as the barcode ‘alphabet’ (Fig. 1a; Supplementary Data). We reasoned that random partial recombination could generate large combinatorial diversity in this substrate. Incubation of the Polylox DNA with Cre recombinase in vitro resulted in incomplete recombination, yielding fragments of five lengths containing nine, seven, five, three or one DNA segments (Fig. 1b) with extensive shuffling (Supplementary Table 1). We targeted the non-expressed Polylox DNA cassette into the Gt(Rosa26)26Sor (also known as Rosa26) locus in embryonic stem (ES) cells (Extended Data Fig. 1a). Rosa26Polylox+/− ES cells were transfected with a plasmid encoding a tamoxifen-inducible version of Cre (MerCreMer) (Methods). Treatment of these cells with 4-hydroxy-tamoxifen (4-OHT) yielded all five DNA fragments after 3 h, followed by increasing appearance of shorter fragments (Extended Data Fig. 1b). In pulse-chase experiments with 4-OHT, loxP site recombinations remained incomplete (Extended Data Fig. 1c), demonstrating that Cre does not strictly act processively.

For in vivo barcoding, we generated Rosa26Polylox+/−C57BL/6 mice. We crossed the Rosa26Polylox allele into mice with ubiquitously expressed, tamoxifen-dependent Cre (Rosa26Polylox/CreERT2) mice and, for endogenous barcoding of HSCs, into mice that express tamoxifen-inducible Cre from the Tek (also known as Tie2) locus14 (Rosa26Polylox/Tie2MCM mice). Barcodes are retrieved from purified cells by single molecule real time (SMRT) sequencing across the entire locus (Extended Data Fig. 1d). Owing to the large sequence differences between the individual DNA

---

Figure 1 | Polylox: a Cre recombinase-driven artificial DNA recombination substrate. a, Structure of the Polylox cassette with loxP sites (triangles; black and white split symbolizes recombination site). Coloured linkers represent DNA segments 1–9. Examples of recombination products resulting from one Cre-mediated excision or one Cre-mediated inversion are shown. The original code segments (‘letters’) are abbreviated 1–9 and their inversions A–I. b, In vitro digestion of Polylox DNA insert in pWP-AG vector by Cre recombinase, and size resolution of recombination products by gel electrophoresis (Supplementary Fig. 1a).

---

1 Division of Cellular Immunology, German Cancer Research Center, D-69120 Heidelberg, Germany. 2 Division of Theoretical Systems Biology, German Cancer Research Center, D-69120 Heidelberg, Germany. 3 Max Delbrück Center, Scientific Genomics Platforms (BIMSB/BH), D-13125 Berlin, Germany. 4 Genomics & Proteomics Core Facilities, German Cancer Research Center, D-69120 Heidelberg, Germany. 5 Bioquant Center, University of Heidelberg, D-69120 Heidelberg, Germany.

*These authors contributed equally to this work.
...underlying barcode generation. The decadic logarithm is used here and throughout the paper. b, Venn diagrams indicating unique and shared barcodes in three independent samples (mice numbers 1, 2 and 3; Extended Data Table 1). c, Generation probabilities of barcodes shared between all mice (n = 177 barcodes; All), shared between two mice (n = 255; 2), or occurring only in one mouse (n = 1,042; Unique) in the Venn diagrams shown in b (red line, median; box ends, 25th and 75th percentiles).

Figure 2 | Combinatorics of Polylox barcoding. a, Illustration of stepwise recombinations, considering only the DNA segments and loxP sites in the red box. The decreasing grey shades indicate an increase in the minimum number of recombination events required to generate a given barcode. b, Calculation of theoretical numbers of barcodes reached with increasing recombination events in the Polylox locus, with a maximum barcode number of more than 1.8 million.

Figure 3 | Polylox barcoding in vivo. a, Barcodes in splenic B cells isolated 18 h after induction of recombination by tamoxifen in a Rosa26PolyloxTie2MCM mouse. Barcodes were ranked according to their generation probability, considering inversions (P_{inv}) equally likely as excisions (P_{exc}) (black line), or half as likely (blue line), or twice as likely (red line). Barcode examples are shown with inversions (blue numbers) and deletions (red numbers) underlying barcode generation. The decadic logarithm is used...
some HSC clones produced oligolineage, myeloid–erythroid output. While these clones underwent multilineage differentiation, we cannot to have originated from single HSC progenitors and hence to represent cytops and EryPs (Fig. 4d). Focusing on the barcodes that were most likely particular clones may sometimes be underestimated. Many embryoni- barcodes overlapped in sample repeats (Fig. 4c). Nevertheless, barcodes to detect such large clones suggests that they are a major ‘structure’ in adult bone marrow.

To analyse barcode propagation from HSCs to peripheral lineages, we sorted the following cells from the mice analysed for adult HSCs (Fig. 4a, b): erythrocyte progenitor (EryP) stages (EryP I–IV) from bone marrow; granulocytes from bone marrow and spleen; B1a, B1b and conventional B (B2) cells of the peritoneal cavity; and B2 cells, CD4 T cells and CD8 T cells from spleen (Extended Data Fig. 5) (1–4 × 10^4 cells per sample). These populations were analysed for barcodes. Many barcodes overlapped in sample repeats (Fig. 4c). Nevertheless, barcodes may be missed in samples, and hence the range of fates acquired by particular clones may sometimes be underestimated. Many embryonically marked HSCs produced outputs5, including short-lived granulocytes and EryPs (Fig. 4d). Focusing on the barcodes that were most likely to have originated from single HSC progenitors and hence to represent HSC clones, we found evidence for multipotent clones (Fig. 4e). While these clones underwent multilineage differentiation, we cannot formally conclude that individual cells within a clone did. However, some HSC clones produced oligolineage, myeloid–erythroid output (CFG4E89, 7456A, E87 and 561). Such lineage-restricted fates suggest that individual cells within an HSC clone show coherent usage of their developmental potential. To gain broader insights into HSC fates in vivo, we considered all barcodes obtained from the peripheral lineages in the two experiments (Fig. 4, Extended Data Fig. 6).

Barcodes that arose in single embryonic HSC progenitors and propagated to the periphery identified productive HSC clones (for further validation of the Pgen threshold, see Extended Data Fig. 3g). To guard against undersampling, we considered only clonal HSC codes that were detected in both repeat samples of at least one lineage. We obtained 117 barcodes, of which nearly half (54) were represented in all assayed lineages, and a further 44 were found in several lineages (Fig. 4f, Extended Data Fig. 6b). Thus the majority of embryonic HSC progenitors give rise to several mature cell types in the adult.

The emergence of barcodes from HSCs into peripheral lineages is also informative regarding lineage pathways. If two peripheral lineages share many barcodes at similar frequencies, they are likely to have arisen from a common developmental pathway; if not, they probably developed more independently. Barcodes found in EryPs and granulocytes were highly concordant (Fig. 5a), whereas EryPs and B2 lymphocytes shared fewer barcodes, and, even when shared, their frequencies were poorly correlated (Fig. 5b; for further cell populations see Extended Data Fig. 7a–c). The Spearman rank correlation coefficient (ρ) quantifies the degree of barcode concordance (Fig. 5c). Clustering all assayed lineages by rank correlation revealed clearly distinct myelo–erythroid and common lymphoid branches (Fig. 5d, Extended Data Fig. 7d). The lack of a clear split between the T and B lymphocyte sub-lineages implies intimate linkage of T and B cell development. The dichotomy of myelo–erythroid and lymphoid
branches was not disturbed when all barcodes were analysed, hence it remained evident even in the presence of non-clonal barcode 'background' (Extended Data Fig. 7e). The two major branches cannot be explained only by lineage-restricted HSC output because distinct myelo–erythroid and lymphoid pathways also emerged from multipotent HSC clones (Extended Data Fig. 7f). Of note, erythrocytes have an approximately 40-fold longer lifespan than granulocytes, but both populations cluster tightly together according to barcode representation. This finding indicates that the different lifetimes of mature lineages do not prevent barcode correlations, and hence that the separation of the common lymphoid branch is not solely due to the longer lifetime of lymphocytes compared to myeloid cells. This notion is further supported by cluster analysis of barcodes from adult HSC clones (Extended Data Fig. 7g). Notably, the B1a sublineage of B cells was a distinct lineage within the common lymphoid branch, clearly separate from B2 cells and T cells; in most but not all cases, the B1b sublineage co-segregated with B1a cells (Fig. 5e, Extended Data Fig. 7e–g).

The original definition of common myeloid progenitors (CMPs) as being restricted to the myelo–erythroid lineage rested largely on transplantation and colony assays, but it is unclear whether CMPs function in this capacity physiologically. Remarkably, all CMP barcodes were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implying that CMP clones are precursors of both lineages. However, about one-third of barcodes in EryPs were found in both EryPs and granulocytes, with highly correlated frequencies (Fig. 5f, g), implies that CMP clones are in vivo precursors of both lineages. However, about one-third of barcodes in EryPs and granulocytes were undetected in CMPs (Fig. 5f, g), raising the possibility of myelo–erythroid differentiation independent of CMPs. CMP barcodes were poorly correlated with lymphoid lineages (Fig. 5h), placing CMPs downstream from the split towards lymphocytes.
barcode comparison places granulocyte–monocyte progenitors (GMPs) also within the myeloid–erythroid branch (Extended Data Fig. 7h).

Finally, we studied output from adult HSCs by treating three Rosa26 Polylox-Tie2\textsuperscript{MCM} mice with tamoxifen at around 8 weeks of age, followed by barcode analysis at 11–13 months of age (Extended Data Fig. 8a). Here, we also included short-term HSCs (ST-HSCs), multipotent progenitors (MPPs), common lymphoid progenitors (CLPs), T and B cell progenitors, and monocyes (Extended Data Fig. 5). Unlike in fluorescent reporter Rosa26 YFP Tie2\textsuperscript{MCM} mice\textsuperscript{14}, barcode induction was not exclusively restricted to HSCs but occurred to a lesser extent also in ST-HSCs, MPPs and CMPs (Extended Data Fig. 8b), probably owing to higher sensitivity of Polylox to Cre recombination versus deletion of the stopper in the Rosa26 YFP locus. The distribution of rare barcodes ($p_{\text{ms}} < 10^{-4}$) again indicated multilineage, oligolineage and unilineage fates (Extended Data Fig. 8c). However, sizable fractions of barcodes in HSCs and progenitors were not found in mature lineages (Extended Data Fig. 8c, d). Sampling repeats showed lower overlap in adult mice (Extended Data Fig. 8e), indicating that clones were smaller overall. While this is compatible with rare output of adult HSCs compared to embryonic HSCs\textsuperscript{14}, the quantitative output of adult HSCs is controversial\textsuperscript{19–21}. Nevertheless, the retrieved barcodes revealed relatedness of granulocyte, monocyte and erythroid lineages on the one hand, and T and B lymphocytes on the other hand (Extended Data Fig. 8f); 3–4 times deeper sampling of four populations (granulocytes, EryPs, B2 cells and CD4 T cells) yielded an even clearer split (Fig. 5i). Likewise, CMPs and GMPs were more closely related to erythrocytes and myeloid lineages than to lymphocytes (Fig. 5j). The paucity of barcodes in CLPs (Extended Data Fig. 8c) precluded robust correlation analysis; however, T and B cell progenitors clustered in their respective lineages, apart from erythroid and myeloid lineages (Fig. 5k).

We have induced Polylox recombination in diverse tissues of ectodermal (brain), mesodermal (muscle, spleen, thymus) and endodermal (liver, lung) origins (Extended Data Fig. 9). Thus, non-invasive Polylox barcoding may aid studies into cellular deconvolution of complex organs and whole organisms at a high resolution. Recently, a Cre-based recombination cassette, similar but not identical to Polylox, was conceived for high-resolution cellular barcoding but has not been tested experimentally\textsuperscript{22}. Polylox is based on the idea that Cre operates incompletely when confronted with many loxP sites in cis under conditions of limited enzyme activity. The barcode diversity offered by Polylox is superior to the diversity of existing fluorescent colour-based endogenous fate-mapping tools\textsuperscript{23, 24}, and, for practical purposes, similar to that of transposon insertion site (TIS) barcoding\textsuperscript{3}; compared to TIS, Polylox has the advantage of a defined genomic location of the barcode (Supplementary Table 3). Our high-resolution fate mapping experiments of embryonic HSC progenitors and adult HSCs both support a bifurcating tree model of haematopoiesis, which was proposed in the 1980s\textsuperscript{24} but has not been tested under physiological conditions. The data imply a haematopoietic architecture with major myelo-erythroid and common lymphoid pathways, while not ruling out additional routes\textsuperscript{25}. Further refinement of precursor–product relationships will require inducible Cre-driven barcoding in progenitor stages downstream of HSCs.
Mice. *Rosa26 Polylox* knock-in mice (B6-G(Rosa)26SorJ-Polylox109J) were generated by gene targeting as described below. For in vivo barcoding experiments, we bred *Rosa26 Polylox* and Tie2CreMcI/− (Tie2CreMcI/−) mice to generate double heterozygous *Rosa26 Polylox* and Tie2CreMcI/− mice (termed *Rosa26 Polylox* and Tie2CreMcI/−), and we bred *Rosa26 Polylox* and Rosa26CreERT2 (B6.129-G(Rosa)26SorJ-Tie2CreMcI/−) mice to generate hemizygous *Rosa26 Polylox* and Tie2CreMcI/− mice. Mice were kept in individually ventilated cages under specific pathogen-free conditions in the animal facility at the German Cancer Research Center (DKFZ). Male and female mice were used, no randomization was done, no blinding was done and no animals were excluded from the analysis. No statistical methods were used to predetermine sample size. For mouse samples sizes, see Extended Data Table 1. All animal experiments were performed in accordance with institutional and governmental regulations, and were approved by the Regierungspräsidium (Karlsruhe, Germany).

**Polylox recombination cassette.** A step-by-step protocol describing the use of Polylox barcoding can be found at Protocol Exchange. We generated a recombination cassette (DNA sequence shown in Supplementary Data), consisting of 3′ and 3′ anchor regions with PCR primer binding sites, flanking nine unique DNA segments, intersected by tenloxP sites with identical recognition sequences but in alternating orientations. The nine segments were chosen from a natural DNA from a plant gene, the cell wall synthase *AT2G1770* gene of *A. thaliana*. To maximize physiological nucleotide distribution, T/A repeats, splice sites and certain restriction sites were removed from the *A. thaliana* AT2G21770 promoter. We inserted the *Polylox* cassette by gene targeting as described below. For the targeted ES cell clone screen, we generated a *Rosa26 Polylox* and Tie2CreMcI/− transfect and isopropanol precipitation. The DNA was amplified by PCR using 1 μg template and the Expand Long Template PCR System (Roche) for 5 min at 95 °C; 30 s at 95 °C, 30 s at 62 °C, 5 min at 72 °C) 35 times; 10 min at the expected size using primer pairs #493 (5′-GCAAGACGCTTTCGACTTGA-3′) and #427 (5′-CATACCTTAGAAGACCTGCTGCAGG-3′) annealing 5′ and 3′ of the *Polylox* cassette, respectively.

**Induction of Polylox recombination in vivo.** For induction of barcodes in adult mice, *Rosa26 Polylox* and Tie2CreMcI/− mice were injected intraperitoneally once, and *Rosa26 Polylox* and Tie2CreMcI/− mice once per day on five consecutive days, with 1 mg tamoxifen or peanut oil for vehicle control. Tamoxifen stock solution was prepared by dissolving 1 g tamoxifen (Sigma) in 4 ml absolute ethanol and 36 ml peanut oil (Sigma) at 55 °C. For *in utero* labelling, timed matings were set up between *Rosa26 Polylox* and Tie2CreMcI/− mice. Nine days after the day of the plug (day 0.5), pregnant mice were treated by oral gavage with a single dose of 2.5 mg tamoxifen to induce barcoding in the developing embryos at E9.5, together with 1.25 mg progesterone (Sigma) to sustain pregnancy. The pups were delivered on E20.5 or caesarean section, raised by foster mothers and genotyped for the Tie2CreMcI allele as described.

**Fluorescence-activated cell sorting (FACS).** Bone marrow (BM) cells were released from the bones by crushing femura, tibiae, pelvis and spine in FACS buffer, and spleens (Sp) were dissociated directly by passing the cells through a 40-μm filter. Peritoneal exudate cells were obtained by lavage of the peritoneal cavity with FACS buffer (PBS supplemented with 5% fetal bovine serum). Cell suspensions were further analysed by Southern blotting using a biotinylated 820-bp probe. The probe was obtained by PCR amplification from genomic C57BL/6 DNA using primers #2,456 (5′-CCATGGGAGATCCACGAAATG-3′) and #2,427 (5′-CCTGGTATCTTATCTTCG-3′). For the indicated times shown in Extended Data Fig. 1b. For the pulse-chase study of 4-OHT induction, MerCreMer-transfected Polylox ES cells were treated with 100 nM 4-OHT for 3 h, then washed, and cultured in 4-OHT-free medium for up to 34 days (Extended Data Fig. 1c). Cell aliquots were collected at the indicated time points, and genomic DNA was purified by phenol–chloroform extraction and isopropanol precipitation. The Polylox cassette was amplified by PCR using 1 μg template and the Expand Long Template PCR System (Roche) for 5 min at 95 °C; 30 s at 95 °C, 30 s at 62 °C, 5 min at 72 °C) 35 times; 10 min at the expected size using primer pairs #493 (5′-GCAAGACGCTTTCGACTTGA-3′) and #427 (5′-CATACCTTAGAAGACCTGCTGCAGG-3′) annealing 5′ and 3′ of the Polylox cassette, respectively.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
lin "Ter119+ CD71+ CD44lo FSCslow")3. For erythroblast isolations, lineage markers included CD3ε, CD11b, CD19, Gr1 and NK1.1. All cell populations were purified on a FACSAriaIII cell sorter (BD Biosciences) using BD FACSDiva software.

Barcode analysis by single molecule real-time (SMRT) sequencing. Genomic DNA from sorted cell populations was purified by phenol–chloroform extraction or proteinase K digestion only. The Polylox cassette was amplified from 100–200 ng template DNA (representing 1.7–3.5 × 10^4 cells) by PCR using primer #2,450 (5′-TGTGGTGATGGCTATTATGATCAG-3′) and primer #2,427 (5′-CATACTTGGAGAAAGCCTGAG-3′), annealing at the 5′ and 3′ anchor regions of the Polylox cassette, respectively, using the following protocol: 5 min at 95 °C; (30 s at 95 °C, 30 s at 56 °C, 5 min at 72 °C) 35 times; 10 min at 72 °C. Second round PCR: 1–2 μl of first PCR reaction was used as template and amplified with primers #2,426 5′-CGACGACACTGCCAAAGATTTC-3′ and #2,427 (see above) for 5 min at 95 °C; (30 s at 95 °C, 30 s at 62 °C, 5 min at 72 °C) 35 times; 10 min at 72 °C. The nested PCR products were purified using the QIAquick PCR Purification Kit (Qiagen), analysed by gel electrophoresis for product length, and analysed by Sanger sequencing (GATC Biotech). Barcodes were decoded for each of these sequences.

Computational analysis of barcode diversity. We iteratively created a library of all barcodes and the adjacency matrix linking the barcodes by all possible excisions and inversions. On this basis, we computed barcode generation probabilities (P_{gen}) using a Markov chain model and factoring in the numbers of Cre recombination events inferred from the experimentally observed barcodes (Supplementary Methods).

Data availability. The computer code used during the current study is available from the corresponding authors on reasonable request. The data sets generated and/or analysed during the current study are available from the corresponding authors on reasonable request.
Extended Data Figure 1 | See next page for caption.
Extended Data Figure 1 | Generation of the Rosa26Polylox locus and experimental procedures for barcode detection and analysis. a, Gene targeting of Polylox DNA into the Rosa26 locus in ES cells; shown are the wild type Rosa26 locus (top), targeting vector (middle) and targeted Rosa26Polylox locus (bottom). Southern blot (insert) (Supplementary Fig. 1b) from genomic tail DNA of control Rosa26+/+ and Rosa26RFP+/+ (ref. 31) mice, and from three Rosa26Polylox+/+ ES cell clones shows restriction fragments corresponding to wild-type (5.8 kb) or targeted (4.8 kb) loci. b, Kinetics of Polylox locus recombination after treatment of Rosa26Polylox MerCreMer ES cells with 4-hydroxy-tamoxifen (4-OHT) at 0 h (Supplementary Fig. 1c). c, Rosa26Polylox MerCreMer ES cells were left untreated and followed for 27 days (left), or pulsed with 4-OHT for 3 h and chased over 34 days (right) (Supplementary Fig. 1d). d, Workflow from cell isolation to Polylox barcode detection. Cell populations of interest were isolated by cell sorting, genomic DNA was purified and the Polylox cassette was amplified by PCR, and the fragments (see Methods) were sequenced by SMRT sequencing using Pacific Biosciences instruments. Raw data were processed with the accompanying software package to obtain the circular consensus sequences (CCS). Subsequently, CCS reads were filtered for reads containing complete Polylox sequences. Next, we aligned the barcode segments to the CCS reads and determined the order and orientation of the segments to retrieve the recombined Polylox barcodes (see Methods). Finally, CCS reads with incomplete segment alignment (X), or illegitimate segment orders (for example, segment duplications) were filtered out and removed from further analysis.
Extended Data Figure 2 | Example of a complex CCS DNA sequence and its corresponding Polylox barcode. a, Schematic drawing of the unrecombined Polylox cassette, and an experimentally found recombined barcode. b, Full nucleotide sequence of one CCS read. In 5’ to 3’ orientation, the DNA sequence is organized into intervening loxP sites and the annotated barcode blocks (‘barcode alphabet’). Numbers and letters refer to the segments shown in a. c, Proportions of unrecombined (blue) and recombined (red) sequence reads in granulocytes (Gr), B cells (B2), CD4 T cells (CD4), and CD8 T cells (CD8) from adult Rosa26PolyloxTie2MCM mice without tamoxifen (TAM) treatment (top) and adult Rosa26Polylox (middle) and adult Rosa26PolyloxTie2MCM (bottom) mice, each treated with tamoxifen as embryos.
Extended Data Figure 3 | Barcode generation probabilities and number of Polylox recombination events in acutely labelled B cells. a, Barcode generation probabilities were computed for a set of barcodes found experimentally (mouse 3, Extended Data Table 1, n = 506 barcodes) with and without length dependence of recombination rate, as described in Supplementary Methods. b, To compare the frequencies of individual barcode segments (‘letters’) generated by the model with experimental data, we focused on data from a Rosa26 Polylox/CreERT2 mouse treated with tamoxifen (Fig. 3a, mouse 1), from which about 15,000 acutely barcoded B cells were analysed. To simulate barcode generation in 15,000 cells, 15,000 barcodes were drawn (with the frequencies of recombination events shown in e). This procedure was repeated 500 times to obtain standard deviations. c, Measured and computed distributions of fragment lengths (experimental data and simulations as in b). d, The observed and measured distributions of the 180 possible pairs of adjacent segments (experimental data and simulations as in b); the unrecombined pairs are particularly abundant. The PacBio instrument loads longer fragments less efficiently than shorter ones. Because of this bias, we restricted the analysis in b–d to fragments with 1, 3 and 5 segments. e, For all barcodes found in B cells from the mouse in b, we computed the minimal number of recombination events (excisions or/and inversions) needed to generate the barcode. All barcodes can be generated with six or fewer recombination events. The cumulative distribution of event frequencies is shown. Similar distributions were obtained in the reported barcoding experiments with Rosa26 Polylox/Tie2MCM mice. f, Barcodes generated once or multiple times in a simulated sample of 15,000 cells (as described in b). On average, 2,920 ± 35 (mean ± s.d.) different barcodes were generated with 15,000 draws. g, Measured barcode frequencies versus computed generation probabilities. For all barcodes retrieved in adult mice after barcode induction in embryonic HSC progenitors (Fig. 4, Exp. 1 and 2), we binned total barcode frequencies according to generation probabilities and calculated boxplot statistics of observed barcode frequencies for each bin (red line, median; box ends, 25% and 75% percentiles; bars, most extreme data points not considered outliers; red dots, outliers, n = 4, 95, 175, 97, 73, 28, 11 barcodes (left) and n = 4, 102, 180, 75, 42, 31, 14 barcodes (right)). From generation probabilities of 0.1 to 10−4, observed median frequency and probability of generation were overall correlated, showing that barcodes generated with higher probability are recovered more frequently. By contrast, for barcodes with generation probabilities <10−4, their median frequency of observation was independent of the probability of generation, indicating that these barcodes have each been generated in the smallest possible unit, a single embryonic HSC progenitor.
Extended Data Figure 4 | Histogram of apparent HSC clone sizes. For the single HSC sequencing data (Fig. 4, Exp. 1 and 2) we show separately the histogram of apparent clone sizes. An apparent clone is defined by all HSCs that contain the same barcode. These apparent clones are unlikely to all be biological clones, owing to the inclusion of abundant barcodes that may have been generated in more than one embryonic HSC progenitor. For the analysis of rare barcodes that are highly likely to define true clones generated from single HSC progenitors, see Fig. 4b, c.
Extended Data Figure 5 | Overview of FACS gating strategies.
a–k, Distinctive surface marker combinations applied for the isolation of specific cell populations are depicted. Pre-gated lineage markers are indicated above the first plot of each panel. Not shown is the additional gating of all populations for size (FSC, SSC) and viability (Sytox blue−). For the complete listing of antibodies and marker phenotypes, see also Methods. a, Isolation of Kit+ Sca1+ stem cells (HSCs and ST-HSCs) and multipotent progenitors (MPPs), upper right, and Kit+ Sca1– myeloid progenitors (CMPs and GMPs), lower right, from bone marrow.
b, Characterization of bone marrow CLPs. c, Definition of pre-B cells (Fr. B and Fr. C) in bone marrow. d, Thymic pre-T cells (DN2 and DN3).
e, Gating of nucleated erythrocyte progenitors in the bone marrow (EryP II–IV, upper right, and EryP I, lower right). f, CD4 or CD8 single-positive T cells from spleen. g, Classical CD19+ splenic B cells. h, Neutrophilic granulocytes from the spleen. i, Splenic monocytes. j, Non-classical B cells (B1a and B1b) from the peritoneal cavity. k, Classical CD19+ B2 cells from peritoneal cavity.
Extended Data Figure 6 | Adult barcode distributions in embryonically induced mice. Mice as in Fig. 4a, Exp. 2. a, Heatmap of all barcodes found in HSCs (first lane) and the indicated erythroid, myeloid and lymphoid lineages in Exp. 2. b, Heatmap of peripheral barcodes ($P_{gen} < 10^{-4}$, and detected in two independent samples of the same population) sorted according to lineage output in Exp. 2. Frequencies of barcodes are represented by colour-coded scales on the right.
Extended Data Figure 7 | See next page for caption.
Extended Data Figure 7 | Clustering of cell types according to all mutual correlations reveals robust dichotomy between common myeloid and common lymphoid development. All data from adult mice with embryonically induced barcodes (Fig. 4a, Exp. 2). a–c, Barcode frequencies in CD8 T cells versus B lymphocytes (B2) (a), in granulocytes (Gr) versus B lymphocytes (B2) (b) and granulocytes (Gr) versus CD4 T cells (c). Data in a–c are from Exp. 1, and each dot is an individual rare barcode with \( n = 48 \) (a), \( n = 49 \) (b) and \( n = 53 \) (c). d, Hierarchical clustering (with distance 1 – Spearman rank correlation coefficient, as described in Fig. 5d) applied to rare and reliably sampled barcodes found in indicated populations in Exp. 2 (\( n = 50 \)). e, Hierarchical clustering as described in d but applied to all barcodes found in peripheral cells in Exp. 1 (\( n = 506 \)) and Exp. 2 (\( n = 496 \)). The inclusion of redundant barcodes reduces differences in correlations, but the split between common myeloid and common lymphoid is evident. f, Clustering as described in d but applied to rare multilineage barcodes (found in at least one myeloid, granulocyte and lymphocyte population, analogous to Fig. 4f; Exp.1, \( n = 16 \) and Exp. 2, \( n = 25 \)). g, Clustering as described in d but applied only to barcodes found in adult HSCs, including redundant ones (shown in Fig. 4d; Exp.1, \( n = 54 \) and Exp. 2, \( n = 56 \)). h, Summary of Spearman rank correlations (mean and 95% confidence bounds computed by non-parametric bootstrap) of GMPs versus the indicated lineages (for CMPs, see Fig. 5h); rare barcodes are from Exp. 2, \( n = 30–44 \).
Extended Data Figure 8 | Polylox barcoding of haematopoiesis in adult mice (all data from adult mice with barcodes induced as adults). a, Barcodes were induced by tamoxifen treatment of adult Rosa26PolyloxTie2MCM mice, and the indicated cell populations were analysed at 11–13 months of age. b, Tamoxifen treatment of adult Rosa26PolyloxTie2MCM mice (Extended Data Table 1) induced Polylox recombination in HSCs and, to a lesser extent, also in downstream stem and progenitor cells, ST-HSCs, MPPs and CMPs (Supplementary Fig. 1c). c, Heatmaps of barcodes satisfying single-cell induction criteria (at the time of labelling) recovered in the indicated stem and progenitor cells, and mature cells in Exp. 3 (left) and Exp. 4 (right). Frequencies of barcodes are represented by colour code (left). d, Heatmaps for individual HSCs satisfying adult single-cell barcode induction criteria, and their lineage output in Exp. 3 (top) and Exp. 4 (bottom). Pgen for the multilineage barcodes were as follows: 1F92G45H3, 1.3 × 10⁻⁹; 123FG45, 2 × 10⁻⁵. e, The barcode overlap between two samples of the same cell population (granulocytes and CD4 T cells isolated from the peripheral blood; 30,000 cells per sample) was smaller than for embryonically labelled mice (Fig. 4c). f, Hierarchical clustering of rank correlations of barcodes from the indicated populations (Exp. 3, n = 129). The colour scale (not shown) for rank correlations is identical to the scale bar shown in Fig. 5i.
Extended Data Figure 9 | Induction of Polylox recombination in tissues of all three germ layers.  

|                | h2o    | Rosa26<sup>+/+</sup> | Rosa26<sup>Polylox</sup> | DNA template                      | PCR result                                      |
|----------------|--------|----------------------|--------------------------|-----------------------------------|------------------------------------------------|
|                | muscle | muscle               | muscle                   | muscle                            | muscle                                         |
|                | thymus | thymus               | thymus                   | thymus                            | thymus                                         |
|                | spleen | spleen               | spleen                   | spleen                            | spleen                                         |
|                | liver  | liver                | liver                    | liver                             | liver                                          |
|                |        | polylox              | polylox                  | polylox                           | polylox                                        |
|                |        | tamoxifen            | tamoxifen                | tamoxifen                         | tamoxifen                                      |
|                |        | oil control          | oil control              | oil control                       | oil control                                    |

Extended Data Figure 9: Induction of Polylox recombination in tissues of all three germ layers.  

To induce Polylox recombination in different tissues in vivo, we crossed the Rosa26<sup>Polylox</sup> allele into mice bearing the Rosa26<sup>CreERT2</sup> allele, which encodes ubiquitously expressed, tamoxifen-regulated Cre, yielding Rosa26<sup>Polylox/CreERT2</sup> mice. 

Adult Rosa26<sup>Polylox/CreERT2</sup> mice were injected with tamoxifen or with oil only (vehicle control) according to the depicted schedule, and were analysed on day 5. 

Genomic DNA was prepared from indicated organs that represent developmental derivatives of all three germ layers: brain (ectoderm), muscle, spleen, and thymus (mesoderm), and liver and lung (endoderm). The Polylox cassette was amplified by PCR, and recombination in each tissue and for all time points was visualized by separating DNA fragments by gel electrophoresis (Supplementary Fig. 1f). The first lane is the PCR water control, the second lane is from Rosa26<sup>+/+</sup> DNA template, and the third lane is from Rosa26<sup>Polylox</sup> (no Cre) template; all other lanes show data from Rosa26<sup>Polylox/CreERT2</sup> mice for the indicated organs and conditions. The DNA sample and PCR result from the muscle oil control were not available.
In utero labelling for embryonic treatment (E9.5, E10.5) was done by oral gavage to the mother. Labelling of adult mice was done by intraperitoneal (i.p.) tamoxifen injection at the age indicated. Time after treatment is given as the time span between first treatment and the day of analysis.

| Figure | Mouse number | Genotype | Sex | Treatment (Age) | Analysis (time after treatment) |
|--------|--------------|----------|-----|-----------------|---------------------------------|
| Fig. 3a | #1 | Rosa26Polylox/CreERT2 f | 1x TAM i.p. (15wks) | 18 hours |
| Fig. 3b, c | #1 | Rosa26Polylox/CreERT2 f | 1x TAM i.p. (15wks) | 18 hours |
| | #2 | Rosa26Polylox/CreERT2 f | 1x TAM i.p. (6mo) | 6 weeks |
| | #3 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 9 months |
| Fig. 4b | #3 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 9 months (Exp. 1) |
| | #6 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 11 months (Exp. 2) |
| Fig. 4c, d | #3 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 9 months (Exp. 1) |
| | #6 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 11 months (Exp. 2) |
| Fig. 4f | #3 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 9 months (Exp. 1) |
| | #6 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 11 months (Exp. 2) |
| Fig. 5a-e | #3 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 9 months (Exp. 1) |
| | #6 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 11 months (Exp. 2) |
| Ext Data Fig. 2c | #4 | Rosa26PolyloxTie2MCM f | No treatment | 7 months |
| | #5 | Rosa26Polylox f | 1x TAM (E10.5) | 8 months |
| | #6 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 11 months |
| Ext Data Fig. 3a | #3 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 9 months (Exp. 1) |
| Ext Data Fig. 3b-e | #1 | Rosa26Polylox/CreERT2 f | 1x TAM i.p. (15wks) | 18 hours |
| | #3 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 9 months (Exp. 1) |
| | #6 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 11 months (Exp. 2) |
| Ext Data Fig. 4 | #3 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 9 months (Exp. 1) |
| | #6 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 11 months (Exp. 2) |
| Ext Data Fig. 5a-h | #8 | Rosa26PolyloxTie2MCM f | 5x TAM i.p. (8wks) | 9 months (Exp. 3) |
| Ext Data Fig. 5j, k | #3 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 9 months (Exp. 1) |
| Ext Data Fig. 6a, b | #3 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 9 months (Exp. 1) |
| Ext Data Fig. 7a-c | #3 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 9 months (Exp. 1) |
| Ext Data Fig. 7d | #6 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 11 months (Exp. 2) |
| Ext Data Fig. 7e-g | #3 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 9 months (Exp. 1) |
| | #6 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 11 months (Exp. 2) |
| Ext Data Fig. 7h | #6 | Rosa26PolyloxTie2MCM m | 1x TAM (E9.5) | 11 months (Exp. 2) |
| Ext Data Fig. 8b | #7 | Rosa26PolyloxTie2MCM f | 5x TAM i.p. (28wks) | 2 weeks |
| Ext Data Fig. 8c, d | #8 | Rosa26PolyloxTie2MCM f | 5x TAM i.p. (8wks) | 9 months (Exp. 3) |
| | #9 | Rosa26PolyloxTie2MCM f | 5x TAM i.p. (8wks) | 11 months (Exp. 4) |
| Ext Data Fig. 8e, f | #8 | Rosa26PolyloxTie2MCM f | 5x TAM i.p. (8wks) | 9 months (Exp. 3) |
| Ext Data Fig. 9c | #11 | Rosa26Polylox/CreERT2 f | 4x oil i.p. | 4 days |
| | #12 | Rosa26Polylox/CreERT2 f | 1x TAM, 3x oil i.p. | 4 days |
| | #13 | Rosa26Polylox/CreERT2 f | 4x TAM i.p. | 4 days |
Experimental design

1. Sample size
Describe how sample size was determined.

The sample size of barcodes resulted from the generation probabilities of each barcode (see Fig. 3a). We considered both all barcodes and those originally generated with high probability in single cells (see text).

2. Data exclusions
Describe any data exclusions.

No data were excluded.

3. Replication
Describe whether the experimental findings were reliably reproduced.

Experimental findings and their replications are described in the manuscript. All findings were reliably reproduced in independent experiments (see text).

4. Randomization
Describe how samples/organisms/participants were allocated into experimental groups.

No randomization was done (see Methods). Barcodes in lineages were compared within individual animals. Hence, samples were collected per animal and not allocated into experimental groups.

5. Blinding
Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blinding was not possible. Cells from single mice were analyzed. Statement that no blinding was done see Methods.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters
For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

| n/a | Confirmed |
|-----|-----------|
| ☑️ | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) |
| ☑️ | A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☑️ | A statement indicating how many times each experiment was replicated |
| ☑️ | The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) |
| ☑️ | A description of any assumptions or corrections, such as an adjustment for multiple comparisons |
| ☑️ | The test results (e.g. p values) given as exact values whenever possible and with confidence intervals noted |
| ☑️ | A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range) |
| ☑️ | Clearly defined error bars |

See the web collection on statistics for biologists for further resources and guidance.
Software
Policy information about availability of computer code

7. Software
Describe the software used to analyze the data in this study.

(i) Mapping of PacBio CCS reads: bowtie2 (version 2.1.0)
(ii) Barcode probability model: Developed customary scripts based on Equations 1-7 in Supplementary Methods, coded in Matlab (Release 2016a).
Detailed description of software provided in Pei, W. et al. Protocol for the use of Polylox – endogenous barcoding for high resolution in vivo lineage tracing. protocol exchange. DOI to be added.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The Nature Methods guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents
Policy information about availability of materials

8. Materials availability
Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

Availability of mice or DNA constructs based on institutional MTAs and subject to developers ongoing scientific work.
9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

| Antibody-Dye (Clone) | Catalog-Number | Lot-Number | Supplier |
|----------------------|----------------|------------|----------|
| CD3e-FITC (17A2)     | 555274         | 69415      | BD Pharmingen |
| CD4-FITC (H129.19)   | 553651         | 66217      | BD Pharmingen |
| CD8a-FITC            | 553030         | 4304688    | BD Pharmingen |
| CD8a-PE (S3-6.7)     | 553032         | 4293593    | BD Pharmingen |
| CD11b-FITC (M1/70)   | 553310         | 24234      | BD Pharmingen |
| CD19-FITC (1D3)      | 553785         | 15995      | BD Pharmingen |
| CD19-APC (1D3)       | 550992         | 66674      | BD Pharmingen |
| CD25-PE-Cy7 (PC61)   | 552880         | 23571      | BD Pharmingen |
| CD43-APC (S7)        | 560663         | 6063818    | BD Pharmingen |
| CD93-BV421 (AA4.1)   | 563806         | 4269682    | BD Pharmingen |
| CD117-APC (2B8)      | 553356         | 3099513    | BD Pharmingen |
| CD11b-PE (M1/70)     | 12011283       | E010741633 | eBiosciences |
| CD11b-PE-Cy7 (M1/70) | 25011282       | 4277098    | eBiosciences |
| CD19-PerCP-Cy5.5 (RA3-6B2) | 4504280 | E083141636 | eBiosciences |
| CD21-biotin (BD9)    | 13021189       | E024501630 | eBiosciences |
| CD23-PE-Cy7 (B38)   | 25023281       | E075311632 | eBiosciences |
| CD24-APC-eFlor780 (M1/69) | 47003282 | E084351637 | eBiosciences |
| CD34-perCP-Cy5.5 (RM4-S) | 45004280 | E082911633 | eBiosciences |
| CD35-APC (S3-7.3)   | 17005181       | E070491630 | eBiosciences |
| CD11b-PE (M1/70)     | 12011283       | E010741633 | eBiosciences |
| CD11b-PE-Cy7 (M1/70) | 25011282       | 4277098    | eBiosciences |
| CD19-PerCP-Cy5.5 (1D3) | 45019382       | E083141636 | eBiosciences |
| CD21-biotin (BD9)    | 13021189       | E024501630 | eBiosciences |
| CD23-PE-Cy7 (B38)   | 25023281       | E075311632 | eBiosciences |
| CD4-eFluor780 (M1/69) | 47004282 | E14514110  | eBiosciences |
| CD3-APC (121-15F9)  | 11589085       | E007151632 | eBiosciences |
| CD4-MHCII-APC (MS/14.15.2) | 16532182 | 4303112  | eBiosciences |
| CD11b-APC-eFlor780 (B28) | 47117182 | 4292975 | eBiosciences |
| CD2-APC (A2F10)     | 17135182       | 4277767    | eBiosciences |
| IgM-FITC (121-15F9) | 11589085       | E007151632 | eBiosciences |
| CD117-APC-eFlor780 (B28) | 47117182 | 4292975 | eBiosciences |
| CD127-PE-Cy7 (A7R34) | 25127182       | E075991635 | eBiosciences |
| CD135-APC (A2F10)   | 17135182       | 4277767    | eBiosciences |
| CD4-Qdot605 (RM4.5) | Q10092         | 1623893    | Life technologies |
| CD4-BV421 (GK1.5)   | 100443         | B213276    | BioLegend |
| CD8a-BV41 (S3-6.7)  | 100753         | B211901    | BioLegend |
| CD11b-BV41 (M1/70)  | 101251         | B212170    | BioLegend |
| CD11c-BV785 (N418)  | 117335         | B201686    | BioLegend |
| CD19-BV421 (6DS)    | 115549         | B212175    | BioLegend |
| CD44-BV605 (IM7)    | 103047         | B207692    | BioLegend |
| CD71-PE-Cy7 (R17217) | 113811        | B138819    | BioLegend |
| CD115-BV605 (AFS98) | 135517         | B209550    | BioLegend |
| CD117-BV711 (2B8)   | 105835         | B210518    | BioLegend |
| CD150-PE           | 115904         | B210719    | BioLegend |
| CD150-BV605 (TC15-12F12.2) | 115927 | B208253 | BioLegend |
| CD150-BV605 (PC136) | 108445         | B206181    | BioLegend |
| CD150-BV605 (PK136) | 108732         | B207544    | BioLegend |
| CD150-BV605 (PC136) | 108445         | B206181    | BioLegend |

For further information on antibodies, see manufacturers’ websites.
10. Eukaryotic cell lines
   a. State the source of each eukaryotic cell line used.

   For gene targeting we used murine embryonic stem (ES) cells (JM8A334 from C57BL/6N, reference Pettitt, S. J. et al. Agouti C57BL/6N embryonic stem cells for mouse genetic resources. Nat Methods 6, 493-495 (2009), and ES cells E14.135 from 129/Ola mice (reference Hooper, M., Hardy, K., Handyside, A., Hunter, S. & Monk, M. HPRT-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured cells. Nature 326, 292-295 (1987)).

   b. Describe the method of cell line authentication used.

   ES cell authentication was achieved via ES cell-derived coat color of chimeric mice and by germ line transmission capability.

   c. Report whether the cell lines were tested for mycoplasma contamination.

   Cell stocks were tested negative for mycoplasma.

   d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

   Not applicable

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

   Provide details on animals and/or animal-derived materials used in the study.

   We generated by gene targeting Rosa26Polylox knock-in mice (B6 Gt(Rosa)26Sortm1(Polylox)Hrr). For in vivo barcoding experiments, we bred Rosa26Polylox/+ and Tie2MCM/+ (Tektm1.1(icre/Esr1*)Hrr) mice to generate double heterozygous Rosa26Polylox/+Tie2MCM/+ mice (termed Rosa26PolyloxTie2MCM). We also bred Rosa26Polylox and Rosa26CreERT2 (B6.129-Gt(Rosa)26Sortm1(cre/ESR1)Tyj/J) mice to generate hemizygous Rosa26Polylox/CreERT2 mice. Male and female mice were used. The age of all experimental animals is given in Exp. Data Tables 1 and 2. All animal experiments were performed in accordance with institutional and governmental regulations, and were approved by the Regierungspräsidium Karlsruhe, Germany.

Policy information about studies involving human research participants

12. Description of human research participants

   Describe the covariate-relevant population characteristics of the human research participants.

   No human studies were done.
Corresponding Author: Hans-Reimer Rodewald
Date: 30 June 2017

Flow Cytometry Reporting Summary

Form fields will expand as needed. Please do not leave fields blank.

- **Data presentation**

  For all flow cytometry data, confirm that:
  
  1. The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
  2. The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
  3. All plots are contour plots with outliers or pseudocolor plots.
  4. A numerical value for number of cells or percentage (with statistics) is provided.

- **Methodological details**

  5. Describe the sample preparation. Bone marrow cells were released from the bones by crashing femura, tibiae, pelvis and spine in FACS buffer, and cells were harvested from the flow-through of a 40-μm filter (BD Falcon). Spleens were dissociated directly by passing the cells through a 40-μm filter or, for monocyte isolation, spleens were dissociated using collagenase D and dispase I. Peritoneal exudate cells were obtained by lavage of the peritoneal cavity with FACS buffer (PBS supplemented with 5% fetal bovine serum).

  6. Identify the instrument used for data collection. BD FACS Aria III Cell Sorter

  7. Describe the software used to collect and analyze the flow cytometry data. BD FACS Diva 6.1.3 Software

  8. Describe the abundance of the relevant cell populations within post-sort fractions. Sorted cell populations had 95-99% purity as determined by post sort re-analysis.

  9. Describe the gating strategy used. The gating strategies are shown in Ext. Data Fig. 4

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information. ❑