METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Mice. The M2/HSB/Tn mice were generated as previously described[11]. To induce transposon mobilization, 8- to 10-week-old male or female mice with the M2/HSB/Tn genotype were fed with 2 mg ml^{-1} Dox together with 5 mg ml^{-1} sucrose in drinking water for 48 h. Thereafter, Dox was removed and successful labelling was verified by retro-orbital sinus peripheral blood collection and analysis (70 μl) after 1 week. All animal procedures were approved by the Boston Children's Hospital Institutional Animal Care and Use Committee. Previous studies have estimated that most haematopoietic lineages are replaced by MPPs within 1–2 months after label[25,29–31]. Thus, for Lin^{+} lineage coupling studies, M2/HSB/Tn mice were analysed within the first 8 weeks after labelling. Since MyPs have limited self-renewal capacity and are rapidly replaced by MPPs, we performed the MyP analysis at short time points after labelling (1 week) and only considered transposon tags not simultaneously present in MPPs.

Bone marrow preparation. After euthanasia, whole bone marrow (excluding the cranium) was immediately isolated in 2% fetal bovine serum in phosphate buffered saline, and erythrocytes were removed with red blood cell lysis buffer. CD45.1 (Ly5.1) mice were used as transplantation recipients (B6.SJL-Ptprca Pep3b/BoyJ, stock 002014, the Jackson Laboratory).

FACS. Lineage depletion was performed using magnetic-assisted cell sorting (Miltenyi Biotec) with anti-biotin magnetic beads and the following biotin-conjugated lineage markers: CD3e, CD19, Gr1, Mac1, and Ter119. Cell populations from bone marrow were purified through four-way sorting using FACSAria (Becton Dickinson) and six-way sorting using MoFlo XDP (Beckman Coulter). The following combinations of cell surface markers were used to define these cell populations. Erythroblasts: 7/4^{-}Ly6G^{-}Ter119^{+}CD71^{+}FSC$^{\mathrm{hi}}$; granulocytes: Ly6G^{+}7-4^{+}B220^{-}Ter119^{-}; monocytes: Ly6G^{-}7/4^{+}B220^{-}Ter119^{-}; pro-/pre-B cells: Ly6G^{-}B220^{+}IL7Ra^{+}; MkP: Lin^{-}Kit^{+}Sca1^{-}CD150^{+}CD41^{+}; MPP1/ST-HSC: Lin^{-}Kit^{+}Sca1^{+}Flt3^{-}CD150^{-}CD48^{-}; MPP2: Lin^{-}Kit^{+}Sca1^{+}Flt3^{-}CD150^{+}CD48^{+}; MPP3: Lin^{-}Kit^{+}Sca1^{+}Flt3^{-}CD150^{-}CD48^{+}; MPP4: Lin^{-}Kit^{+}Sca1^{+}Flt3^{+}CD48^{+}; LT-HSC: Lin^{-}Kit^{+}Sca1^{+}Flt3^{-}CD150^{+}CD48^{-} (±CD41). Other populations are defined in Supplementary Table 1. Representative examples of sorted populations are shown in Supplementary Figs 1–3. Flow cytometry data were analysed with FlowJo (Tree Star). For transposon tag content extraction and analysis, we FACS-sorted all the available cells from the whole bone marrow extract using purity modes (approximately 98% purity) at about 75–80% efficiency. The antibodies (their clone number, the commercial house, and concentration) were as follows: Ly6B.2 FITC (7/4, Miltenyi, 1:100), Ly6G Alexa Fluor 700 (1A8, eBiosciences, 1:50), Ter119 APC (TER119, eBiosciences, 1:100), CD71 BV510 (C2, BD biosciences, 1:100), CD45R(B220) eFluor 450 (RA3-6B2, eBiosciences, 1:100), CD19 APC/Cy7 (1D3, eBiosciences, 1:50), CD127(IL-7Rα) PE/Cy7 (A7R34, Biolegend, 1:25), CD117 (Kit) FITC/APC (2B8, eBiosciences, 1:100), Ly6a (Sca1) PE/Cy7 (D7, eBiosciences, 1:100), CD135 (Flt3) APC (A2F10, Biolegend, 1:25), CD150 PE/Cy5 (TC15-12F12.2, Biolegend, 1:100), CD48 APC/Cy7 (HM48-1, BD biosciences, 1:100), CD41 BV605 (MwReg30, Biolegend, 1:100), CD3e biotin (145-2C11, eBiosciences, 1:100), CD19 biotin (MB19-1, eBiosciences, 1:100), Gr1 biotin (RB6-685, eBiosciences, 1:100), CD11b (Mac1) biotin (M1/70, eBiosciences, 1:100), Ter119 biotin (TER119, eBiosciences, 1:100), streptavidin eFluor 450 (eBiosciences, 1:200), FcgRII/III eFluor 450 (93, eBiosciences, 1:100), CD34-FITC (RAM, eBiosciences, 1:25), CD42 APC (HIP1, Biolegend, 1:100), CD9 PE (MZ3, Biolegend, 1:200).

Transplantation assays. Whole bone marrow cells or sort-purified LT-HSCs from M2/HSB/Tn mice were transplanted in 150 μl of αMEM (Gibco, Thermo Fisher Scientific) through retro-orbital injection into γ-irradiated recipient mice (split dose of 2.5 + 2.5 Gy for sublethal irradiation, and 5.5 + 5.5 Gy for lethal irradiation, with a 2 h interval). Donor cell engraftment and label frequency were analysed after 16 weeks using LSRII equipment (Becton Dickinson).

HSC culture assays. One thousand sort-purified LT-HSCs from M2/HSB/Tn mice were cultured together with 10,000 MS-5 stromal cells in round-bottom 96-well plates together with SCF (100 ng ml^{-1}), TPO (100 ng ml^{-1}), Flt3L (50 ng ml^{-1}), IL7 (20 ng ml^{-1}), IL3 (10 ng ml^{-1}), IL11 (50 ng ml^{-1}), and GM-CSF (20 ng ml^{-1}) in αMEM with 1% penicillin/streptomycin and 10% FCS (Thermo Fisher) for 2 weeks, changing the medium 24 h after sort and then every 48 h (Becton Dickinson). Myeloid and lymphoid HSC progeny was FACS-sorted after labelling with Gr-1, Mac-1, CD19, and B220 antibodies (eBiosciences). All growth factors and cytokines were mouse recombinant and purchased from Peprotech.

DNA isolation and amplification. Cells of interest were sorted into 1.7-ml tubes and concentrated into 5–10 μl of buffer by low-speed centrifugation (700*g* for 5 min). Samples with fewer than 10,000 cells were subjected to whole-genome amplification with a Phi29 kit (Epicentre/Lucigen) according to the manufacturer's instructions. Samples with more than 10,000 cells were purified by a QIAamp DNA Micro kit (56304, Qiagen).

TARIS. Our original technique for molecular identification of transposon integration sites was based on ligation-mediated (LM) PCRs. Others and we have observed significant tag amplification biases with this method, which limit the quantitative potential of the clonal data obtained[11,32,33]. To improve the current technique, we developed a method based on TARIS (Extended Data Fig. 1). This method provided similar sensitivity levels as LM-PCR but more quantitatively and reproducibly captures the clonal composition of complex samples (Extended Data Fig. 2). For TARIS, the total purified DNA was subjected to enzymatic restriction with 10 U of HindIII-HF (NEB) overnight. TARIS adaptor primer was hybridized and extended using 1 U Klenow DNA polymerase (NEB) for 2 h. Then, total DNA was cleaned up using AMPure XP SPRI beads (Beckman Coulter) and used as a template for a 20 μl T7 RNA polymerization reaction (NEB, High Yield Hiscribe T7 kit) overnight. Then, the template was digested with 1 U of Turbo DNase (Ambion) and the RNA product was polyadenylated using 1 U of polyA RNA polymerase (NEB). The polyA RNA was purified with SPRI beads, and then converted into cDNA using iScript reverse transcriptase (Biorad). TARIS cDNA was used as template for 30 PCR cycles using the HSB-transposon-specific Tn-1C, the MAF-Tn-1F, and the MAR-polyT primers for 30 cycles, and then 12 cycles of indexing PCR using the MP1 and ID primers (ID1-48) and the KAPA HiFi PCR kit. Solexa sequencing was performed on HiSeq 2000 (Illumina) at the Tufts Genomics Core. Tag identification and alignment was performed as previously described[11]. In brief, we extracted the transposon-containing reads from each fastq file, trimmed the adaptor and transposon sequences, and aligned the integration sites to the reference mouse genome (Ensembl mm9) using Bowtie 1.2. Then, reads were normalized between samples (per million reads). Sequences were always compared with at least one additional independently labelled mouse, with libraries prepared in parallel and sequenced in the same HiSeq lane to account for contaminations. Tags present in the control mouse samples were filtered out (contaminating reads). Next, read frequencies were column-normalized, and graphs were coloured using a logarithmic scale. For hierarchical clustering based on transposon tag distribution, we first determined the Spearman's correlation matrix for the compared populations and then performed agglomerative clustering (single method) using (1 − correlation coefficient) as the distance metric. Curve fitting was performed with the Lowess function. All indicated statistical tests were two-tailed parametric *t*-tests using Welch's s.d. correction (exceptions are mentioned where appropriate). Data visualization and statistical analysis was performed using Microsoft Excel, R (version 3.3.1), and GraphPad Prism (version 7). Primers used were TARIS adaptor primer (5′-GCATTAGCGGCCGCGAAATTAATACGACTCACTATAGGGAGTCTAAAGCCATGACATC-3′), Tn1-C primer (5′-CTTGTGTCATGCACAAAGTAGATGTCC-3′), MAF-Tn1-1F primer (5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCGAGTTTTAATGACTCCAACT-3′), and MAR-polyT primer (5′-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTV-3′). All primers were ordered from IDT DNA technologies, at 100 nmole scale and HPLC-purified.

Single-cell RNA sequencing and low-level data processing. Transcriptome barcoding and preparation of libraries for single-cell mRNA sequencing was performed using the most up-to-date inDrops protocol[34]. For our experiment, the Lin-Sca1^{+}Kit^{+} bone marrow fraction from a single BL6 mouse was labelled and FACS-sorted to purify the entire LT-HSC, MPP1, MPP2, MPP3, and MPP4 fractions. Approximately 2,000 cells of each fraction were encapsulated and libraries for all the populations were prepared the same day, with the same stock of primer-gels and RT-mix. Libraries were sequenced on an Illumina NextSeq 500 sequencer using a NextSeq High 75 cycle kit: 35 cycles for read 1, 6 cycles for index i7 read, and 51 cycles for read 2. Raw sequencing reads were processed using the inDrop pipeline previously described, with the following modifications: Bowtie version 1.1.1 was used with parameter –e 100; all ambiguously mapped reads were excluded from analysis; and reads were aligned to the Ensembl release 81 mouse mm10 cDNA reference.

Data visualization using SPRING. We combined mRNA count matrices from five simultaneously processed and indexed libraries (LTHSC-2A, STHSC-2A, MPP4-2A, MPP3-2A, MPP2-2A). Cells with few mRNA counts (<1,000 unique molecular identifiers) and stressed cells (mitochondrial gene-set Z-score > 1) were filtered out[35]. The remaining high-quality cells (4,248) were total-counts normalized. We next filtered genes, keeping those that were well detected (mean expression > 0.05) and highly variable (CV > 2). Finally, we reduced dimensionality by Z-scoring each gene and applying principal component analysis (PCA), retaining the top 50 principal components. The cells were then visualized using SPRING, a graph-based single-cell viewing interface[36]. Visual inspection of the SPRING plot revealed a strong cell-cycle signature defined by high expression of genes associated with the G2/M phase (*Ccnb1*, *Plk1*, *Cdc20*, *Aurka*, *Cenpf*, *Cenpa*, *Ccnb2*, *Birc5*, *Bub1*, *Bub1b*, *Ccna2*, *Cks2*, *E2f5*, *Cdkn2d*). Hypothesizing that this