

Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis

<https://doi.org/10.1038/s41586-020-2503-6>

Received: 13 June 2019

Accepted: 28 April 2020

Published online: 15 July 2020

 Check for updates

Alejo E. Rodriguez-Fraticelli^{1,2,3,4}, Caleb Weinreb⁵, Shou-Wen Wang⁵, Rosa P. Migueles⁶, Maja Jankovic^{1,2}, Marc Usart^{1,2}, Alton M. Klein⁵, Sally Lowell⁶ & Fernando D. Camargo^{1,2,3,4} 

Bone marrow transplantation therapy relies on the life-long regenerative capacity of haematopoietic stem cells (HSCs)^{1,2}. HSCs present a complex variety of regenerative behaviours at the clonal level, but the mechanisms underlying this diversity are still undetermined^{3–11}. Recent advances in single-cell RNA sequencing have revealed transcriptional differences among HSCs, providing a possible explanation for their functional heterogeneity^{12–17}. However, the destructive nature of sequencing assays prevents simultaneous observation of stem cell state and function. To solve this challenge, we implemented expressible lentiviral barcoding, which enabled simultaneous analysis of lineages and transcriptomes from single adult HSCs and their clonal trajectories during long-term bone marrow reconstitution. Analysis of differential gene expression between clones with distinct behaviour revealed an intrinsic molecular signature that characterizes functional long-term repopulating HSCs. Probing this signature through in vivo CRISPR screening, we found the transcription factor TCF15 to be required and sufficient to drive HSC quiescence and long-term self-renewal. *In situ*, *Tcf15* expression labels the most primitive subset of true multipotent HSCs. In conclusion, our work elucidates clone-intrinsic molecular programmes associated with functional stem cell heterogeneity and identifies a mechanism for the maintenance of the self-renewing HSC state.

To simultaneously analyse mRNA and lineage information for multiple stem cell clones, we isolated long-term HSCs (LT-HSCs) from 8-week-old mice and transduced them with the lineage and RNA recovery (LARRY) lentiviral barcoding library¹⁸ (Fig. 1a). We transplanted approximately 1,000 labelled cells into lethally irradiated 8-week-old recipients and analysed the HSC and committed progenitor cell fractions by inDrop single-cell RNA-seq (scRNA-seq) after steady-state repopulation at 16–24 weeks after transplant (Extended Data Fig. 1a; $n = 3$ experiments (5 mice)). We used Louvain clustering to identify different stem/progenitor populations, and these were labelled and merged on the basis of expression of previously identified markers (Extended Data Fig. 1b, c, Supplementary Table 1). We then assigned LARRY lentiviral barcodes to each cell to reconstruct clonal relationships. Importantly, we benchmarked LARRY for long-term clonal tracking, confirming that library diversity was adequate for single-cell labelling, that barcode calling was efficient for most populations, that single-cell readouts accurately represented the DNA barcodes, and that barcode silencing was negligible (Extended Data Fig. 1d–m).

Evaluation of HSC and progenitor barcodes confirmed that transplantation haematopoiesis is sustained predominantly by HSCs, with most progeny represented in at least one barcoded HSC, as previously suggested^{3,19–21} (Extended Data Fig. 2a). This experimental framework allowed us to analyse the functional behaviours of 227 HSCs and their

associated gene expression programmes. We observed a large degree of clonal heterogeneity in terms of progeny output activity (A_i), which is defined as the ratio between the abundance of a given clone i in the committed progenitor pool and its frequency in the HSC compartment (range: 0–51, mean = 1.66; Fig. 1b, c, Extended Data Fig. 2b, c). Remarkably, over 55% of HSC clones (~60% of all HSCs) were categorized as relatively ‘low output’, self-renewing significantly more than differentiating ($A_i < 1$; Fig. 1d, Extended Data Fig. 2d). Importantly, detection of low-output clones was not an artefact of sampling, as clones containing as many as 588 cells showed this behaviour (Extended Data Fig. 2b, c). While previous retroviral barcoding studies have suggested the existence of low-output clones, our single-cell approach allowed us to precisely quantify and appreciate the extent of this behaviour^{8,11,19,22–25}. We also found that HSC clones were highly diverse in their lineage bias (B_i), which is defined as the frequency ratio between any single lineage and the other progenitors. In particular, we found that ~30% of clones presented megakaryocyte (Mk)-biased output and were responsible for 50–60% of all Mk progeny (Fig. 1e–g, Extended Data Fig. 2e), in line with previous observations^{8–10}.

In addition to defining clonal HSC heterogeneity, our approach simultaneously allowed us to characterize differences in gene expression among functionally different clones. Compared to high-output HSCs, low-output HSC clones expressed higher levels of quiescence

¹Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. ²Boston Children’s Hospital Stem Cell Program and Department of Hematology/Oncology, Boston, MA, USA. ³Harvard Stem Cell Institute, Cambridge, MA, USA. ⁴Harvard Medical School, Department of Pediatrics, Boston, MA, USA. ⁵Harvard Medical School, Department of Systems Biology, Boston, MA, USA. ⁶Institute for Stem Cell Research, MRC Centre for Regenerative Medicine, The University of Edinburgh, Edinburgh, UK.
✉ e-mail: fernando.camargo@childrens.harvard.edu

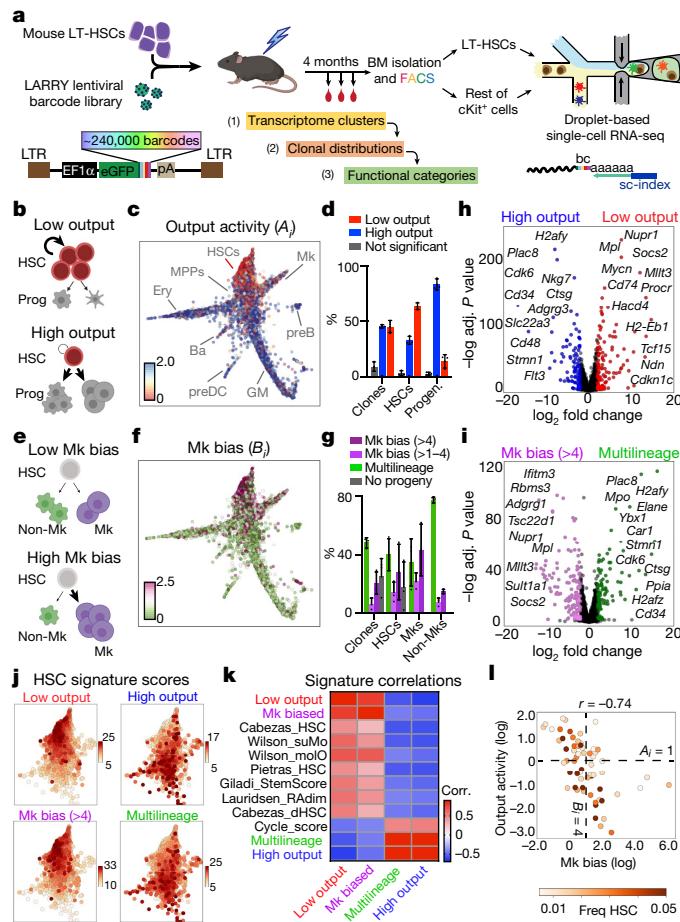


Fig. 1 | Simultaneous single-cell lineage and transcriptome sequencing maps functional HSC heterogeneity. **a**, Experimental design for studying HSC heterogeneity with the LARRY lentiviral barcoding library. All panels are representative from $n=3$ independent labelling experiments (five mice). LTR, long terminal repeat; bc, barcode. **b**, Schemes of low-output (top) and high-output (bottom) HSC clones. Prog, progenitors. **c**, Single-cell map that shows clonal HSC output activity values. LARRY barcodes are used to assign each cell to an HSC clone, and then all cells from each clone are coloured based on its calculated output activity. Major progenitor cell populations are labelled. Ba, basophil; Ery, erythroid; GM, granulocyte–monocyte; MPP, multipotent progenitor; preB, pre-B cell; preDC, pre-dendritic cell. **d**, The distribution of high-output (output activity of >1) and low-output (output activity of <1) HSC cells and clones (shown as % of total HSCs). Mean \pm s.d. is shown. **e**, Schemes of lineage-balanced (top) and lineage-biased (bottom) HSC clones. **f**, Single-cell map showing clonal Mk-bias values. **g**, Distribution of Mk-biased and multilineage HSCs (cells and clones), Mk cells and non-Mk cells (shown as % of total). Mean \pm s.d. is shown. **h**, Genes differentially expressed in low-output (right, $n=7,254$ cells) versus high-output (left, $n=3,512$ cells) HSCs. Genes with an adjusted $P < 0.01$ (Benjamini–Hochberg-corrected t -test) and fold change of >2 are coloured. Selected genes are labelled. **i**, Genes differentially expressed in Mk-biased (right, $n=3,399$ cells) versus multilineage (left, $n=3,771$ cells) HSCs. Genes with an adjusted $P < 0.01$ (Benjamini–Hochberg-corrected t -test) and fold change of >2 are coloured. **j**, Single-cell map of HSCs, coloured by the signature score values. **k**, Heatmap that shows the Pearson correlation between different signature scores across all HSCs ($n=10,837$). **l**, Scatter plot of Mk bias and output activity (log-transformed) for each HSC clone, coloured by clone HSC frequency (freq). The dashed lines are the output activity threshold ($A_i=1$) and the Mk-bias threshold ($B_i=4$). Only clones with a HSC frequency of >0.005 are depicted ($n=62$).

and self-renewal markers such as *Txnip*, *Mllt3*, *Socs2*, *Mpl*, *Mycn*, *Cdkn1c* and *Ndn*, in addition to other components poorly described in HSCs, including those encoding fatty-acid oxidation enzymes (*Hacd4*), MHC

class II components (*Cd74* and *H2-Eb1*) and transcription regulators (*Nupr1* and *Tcf15*)^{26–32} (Fig. 1h). Interestingly, the low-output HSC signature shared multiple genes with the Mk-biased HSC signature (Fig. 1i, Supplementary Table 2). Analysis of computed signature scores confirmed that the low-output and Mk-biased genes are co-expressed and overlap with published signatures of highly purified native LT-HSCs, while they negatively correlate with the cell-cycle signature score, suggesting a relatively quiescent HSC state post-transplantation^{12,17,33–35} (Fig. 1j, k, Extended Data Fig. 2f). The barcode measurements of HSC output activity (A_i) and Mk bias (B_i) also presented a significant negative correlation ($r=-0.74$), confirming that low-output and Mk-biased behaviours are enriched in the same set of clones ($P < 0.001$; Fig. 1l). Importantly, these behaviours were not restricted to distinct HSC subpopulations defined solely by transcriptional clustering methods, highlighting the relevance of clonal tracking for studying HSC heterogeneity (Extended Data Fig. 3a–e, Supplementary Table 3).

Altogether, our data suggest that, even after transplantation, a large number of engrafted HSC clones display low progeny output (irrespective of their clone size), contribute biasedly to the Mk lineage, and express a distinct signature that contains genes associated with increased quiescence and self-renewal. We posit that, after transplantation, a subset of HSCs reacquire a configuration that resembles unperturbed native LT-HSCs, which also contribute little to mature progeny during the first year of life⁵ and show predominant Mk-lineage contribution^{8,9,36}.

The genetic programme of HSC engraftment

To identify clone-intrinsic gene expression programmes associated with functional long-term repopulation capacity, we performed secondary transplants. We repeated our barcoding experiments, sampling only half of the LT-HSC compartment by inDrop at 16 weeks (1T clones), while the other half of the HSCs (~3,500 barcoded cells) was randomly split into two equal parts and transplanted into two secondary recipients. These recipients were analysed 24 weeks after transplantation by inDrop (2T clones; 25,636 cells) (Fig. 2a, b). We found a strong correlation ($r=0.67$) between the secondary engraftment potential ('2T-expansion') of the same clones in separate secondary recipients (Fig. 2c, Extended Data Fig. 4a), in line with a recent report³⁷. This high correlation seems to be predetermined, at least in part, by size-independent clone-autonomous properties of the primary HSC clone (Fig. 2c, Extended Data Fig. 4b, c), when compared to an equipotent null model, in which each HSC is assumed to have equal probability of engrafting ($P=0.0013$; see Supplementary Methods). Surprisingly, we found that at least one of the properties affecting secondary engraftment was the HSC output activity in the primary recipient. We found that high-output 1T HSC clones were significantly absent in secondary recipients (Fig. 2d, e). Instead, serial transplantation was mainly driven by low-output 1T HSC clones ($P=0.049$, compared with the equipotent null model; Fig. 2d–f, Extended Data Fig. 4d), and this observation held true when considering each lineage separately or all combined (Mk, myeloid or lymphoid; Extended Data Fig. 4e). Together, these results argue that the differentiation history of a stem cell clone compromises its long-term repopulating capacity in a clone-autonomous manner.

Similar to other clonal functional outcomes, serial repopulating behaviour was only modestly enriched in HSC subclusters defined solely by their transcriptome (Extended Data Fig. 4f). To extract a gene signature that was indicative of long-term potential, irrespective of clustering or any other parameters, we performed differential expression analysis that compared clones with observed serial repopulation and clones that were not detected in the second grafts. The molecular signature of functional long-term regeneration was characterized by expression of several well-known markers of native quiescent HSCs (*Mycn*, *Procr*, *Mllt3*, *Matn4*, *Hoxb8*, *Slamf1*, *Rorc* and *Cdkn1c*)^{7,16,31,32,38–41}, and by lack of expression of cycling/activated HSC and Mk-differentiation markers

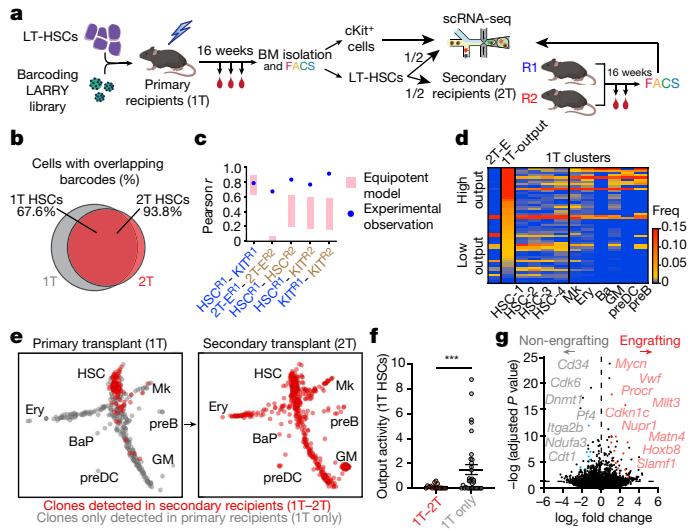


Fig. 2 | A clonal molecular signature of serial repopulation capacity.

a, Experimental design for the secondary transplantation experiment. **b**, Venn diagram that shows the clonal overlap between 1T HSCs (% of cells) and 2T HSCs. **c**, Histogram of Pearson correlations between secondary recipient clone measurements (see Supplementary Methods). The pink bars show the correlation distribution of the equipotent HSC null model (1 s.d. over 10^4 calculations). The blue circles represent the observed experimental data. 2T-E, clone expansion in each secondary recipient. **d**, Heatmap that shows the clonal frequency in 2T and in 1T clusters. The clones are ordered from top to bottom by 1T output activity (the scale is normalized to the plot with the same scale). Only clones represented in at least five 1T HSCs are shown. **e**, SPRING plot of

clones in 1T (left) and clones in 2T (right), randomly subsampled for visualization (representative from $n=2$ animals). Clones are coloured red if they are also detected in 2T (1T-2T clones) and in grey if they are not detected in 2T (1T only). Populations are labelled. **f**, Scatter plot that shows the output activity (A_1) of 1T HSC clones comparing 2T-engrafting (red, $n=17$) to non-engrafting (grey, $n=33$) clones. Lines represent mean \pm s.e.m. *** $P=0.0098$ using the Kolmogorov–Smirnov (two-sided) test. **g**, Volcano plot of differential expression analysis of secondary engrafting ($n=773$) versus non-engrafting ($n=591$) HSCs. Benjamini–Hochberg-corrected t -test P values are shown. The horizontal dashed line indicates $P=0.05$.

(*Cd34*, *Cdk6*, *Pf4*, *Itga2b* and *Gata1*)^{16,42–44}, in addition to a large number of genes that are yet undefined in this process (Fig. 2g, Supplementary Table 2). This signature correlated remarkably with the low-output and Mk-biased signatures, and with some native LT-HSC signatures that have been previously described^{12,17} (Extended Data Fig. 5a–c, Supplementary Table 2). Altogether, our results indicate that long-term potency is an intrinsic and heritable property of self-renewing low-output HSC clones that can propagate through transplantation and is characterized by the maintenance of a unique transcriptional programme, which resembles the programme of native and quiescent HSCs.

In situ CRISPR screening of HSC fate

On the basis of the combined transcriptional signatures of low-output and secondary-repopulating HSC clones, we selected 63 differentially upregulated genes previously uncharacterized in HSCs, to test their requirement for suppressing HSC output (Supplementary Table 4). We performed a doxycycline (Dox)-inducible positive-enrichment in vivo CRISPR screening post-reconstitution to identify single guide RNAs (sgRNAs) that increased HSC contribution to mature/progenitor cell fractions^{45,46} (Fig. 3a, Supplementary Table 5). Deep sequencing revealed five targets that were consistently overrepresented in most populations and had the highest positive average enrichment score using MAGeCK analysis: *Adam22*, *Tcf15*, *Clec2d*, *Clca3a1* and *Smtnlu*⁴⁷ (Fig. 3b, Supplementary Table 6). We determined that *Tcf15* sgRNA had the most robust effect across the six biological replicates (Extended Data Fig. 7a). Moreover, TCF15 (also known as Paraxis) is a known transcription factor, which suggested a possible master regulatory function in the molecular programme that controls HSC output. TCF15 is essential for pluripotency exit, somitogenesis and paraxial mesoderm development, but has not been described in haematopoiesis so far^{48–50}.

We confirmed that *Tcf15* expression is specific to HSCs in our and previously published data sets^{9,14,51} (Extended Data Fig. 6a–c). *Tcf15*

expression correlated with low-output/long-term engraftment HSC signatures (Extended Data Fig. 6d–g). Clonal data showed that *Tcf15*^{hi} HSCs exhibited significantly lower output activity (Extended Data Fig. 6h, i). In addition, combined single-cell mRNA and sgRNA sequencing revealed that *Tcf15* sgRNA clones were partially depleted from transcriptionally defined quiescent HSC clusters and enriched in committed progenitor clusters (Fig. 3c, Extended Data Fig. 7b). Differential gene expression analysis in *Tcf15* sgRNA cells showed reduced expression of *Tcf15* (expression: 13% of control, $P=0.02$), in addition to other quiescent HSC markers (*Sult1a1*, *Procr*, *Mecom* and *Cdkn1b/c*), and concomitant upregulation of cell-cycle and active HSC hallmarks (Fig. 3d, Supplementary Table 7).

A typical consequence of loss of quiescence is stem cell exhaustion and impaired long-term regenerative capacity^{52,53}. Lentiviral-mediated *Tcf15* CRISPR knockout partially impaired peripheral blood and bone marrow engraftment in primary transplants (Extended Data Fig. 7c–f). The most noticeable defect was observed in the immunophenotypic LT-HSC gate, suggesting a specific loss of the most quiescent stem cells, which we confirmed by cell-cycle analysis (Fig. 3e, f, Extended Data Fig. 7g). We further validated that disrupting *Tcf15* fully abrogates long-term engraftment potential in secondary transplantation (Fig. 3g).

Since TCF15 is a transcription factor, we hypothesized that inducing *Tcf15* expression could be sufficient to enforce quiescence through the upregulation of a *Tcf15*-driven gene network. Using a lentiviral Dox-inducible *Tcf15* transgene, we first observed that *Tcf15* overexpression inhibited HSC proliferation in vitro (Extended Data Fig. 8a, b). Similarly, *Tcf15* overexpression in stably reconstituted mice led to the inhibition of haematopoietic differentiation (Extended Data Fig. 8c, d). Remarkably, *Tcf15*-overexpressing cells exhibited a 20.8-fold enrichment in the frequency of LT-HSCs in the bone marrow and a depletion of downstream progenitors (Extended Data Fig. 8e–i). scRNA-seq analysis of the cKit⁺ marrow fraction revealed that *Tcf15*-overexpressing cells were almost exclusively restricted to

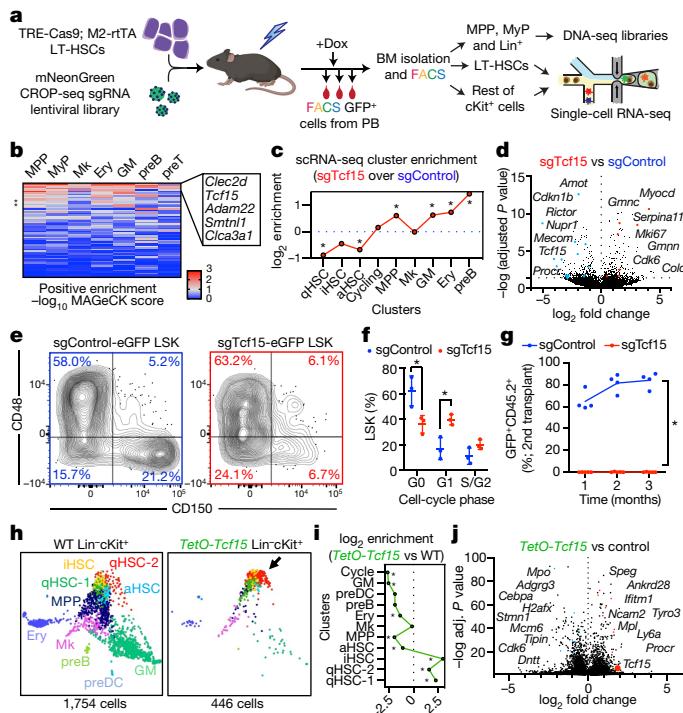


Fig. 3 | In vivo CRISPR screening identifies regulators of HSC output.

a, Experimental design for the steady-state CRISPR screening. MyP, myeloid progenitors; PB, peripheral blood. **b**, Heatmap that shows the positive enrichment score for each targeted gene (rows), in each bone marrow (BM) compartment (columns). The top five genes are labelled. preT, pre-T cell. The asterisks indicate putative positive controls. **c**, Single-cell cluster enrichment of the *Tcf15* sgRNA (sgTcf15; log₂ fold over sgControl). * $P < 0.1$ by differential proportion analysis (DPA) test ($n_{\text{sgTcf15}} = 298$, $n_{\text{sgControl}} = 437$). For DPA, see Methods. aHSC, activated HSC; iHSC, HSC with inflammation signature; qHSC, quiescent HSC. **d**, Volcano plot that shows differentially expressed genes comparing sgTcf15 ($n = 220$) to sgControl ($n = 269$) HSCs from the scRNA-seq experiments. Benjaminini–Hochberg-corrected *t*-test *P* values are shown. **e**, Fluorescence-activated cell sorting (FACS) plots that show bone marrow LSK staining for SLAM in donor-derived sgControl and sgTcf15 eGFP⁺ cells. Plots are representative from $n = 4$ independent experiments. **f**, Quantification of the cell-cycle status of eGFP⁺ LSKs. Mean \pm s.d. is shown. * $P < 0.005$ ($n = 3$, Holm–Sidak-corrected two-sided *t*-test). **g**, Quantification of donor engraftment (% eGFP⁺ of all peripheral blood cells) in secondary transplantation. * $P < 0.005$ ($n = 4$, Holm–Sidak-corrected two-sided *t*-test). **h**, SPRING scRNA-seq map of one representative experiment comparing wild-type (WT; left) to *Tcf15*-overexpressing cKit-enriched cells (right). **i**, Cluster enrichment of *TetO-Tcf15* represented as a log₂ fold-enrichment over control. * $P < 0.1$ DPA test ($n_{\text{TetO-Tcf15}} = 440$, $n_{\text{control}} = 1,752$). **j**, Volcano plot showing differential gene expression of *TetO-Tcf15* ($n = 446$) versus control cKit⁺ ($n = 1,754$) cells. Benjaminini–Hochberg-corrected *t*-test *P* values are shown.

the quiescent HSC clusters (Fig. 3h, i). Secondary transplants demonstrated that *Tcf15*-overexpressing LT-HSCs could still exhibit long-term repopulation upon suppression of *Tcf15* transgene expression by Dox withdrawal (Extended Data Fig. 8j, k). To outline a gene programme driven by TCF15, we compared the single-cell differential gene expression signatures of *Tcf15*-overexpressing (Fig. 3j, Supplementary Table 8) and *Tcf15*-depleted HSCs, and found 174 genes with significant symmetrically opposite expression, which were enriched for previously described regulators of HSC quiescence/maintenance, including *Cdkn1c*, *Socs2*, *Mcl1* and *Gata2*^{29,31,54,55} (Supplementary Table 9). Altogether, these experiments indicate that *Tcf15* expression is both required and sufficient to maintain stem cell quiescence, and that TCF15 is required for the long-term regenerative capacity of HSCs.

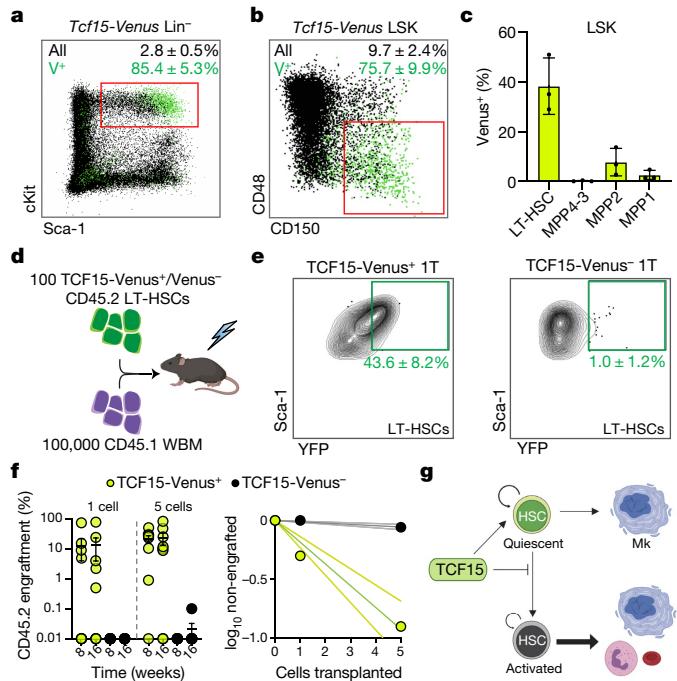


Fig. 4 | TCF15 expression defines the functional LT-HSCs. **a**, FACS plot of the *Tcf15*-Venus knock-in reporter Lin[−] cells. Mean \pm s.d. % of LSKs (red square) of all Lin[−] (Venus⁺ (V) versus all cells) is shown. Plots in **a**, **b** and **e** are representative from $n = 3$ independent experiments with similar results. **b**, FACS plot of *Tcf15*-Venus knock-in reporter LSK Venus⁺ cells stained for SLAM markers. Mean \pm s.d. % of LT-HSCs (red square) within LSK (Venus⁺ versus all cells) is shown. **c**, Mean \pm s.d. percentage of *Tcf15*-Venus expression within each LSK SLAM compartment ($n = 3$). **d**, Primary competitive transplantation of HSCs derived from *Tcf15*-Venus reporter (CD45.2) mice. WBM, whole bone marrow. **e**, FACS plots that show YFP (Venus) versus Sca-1 intensity of donor-derived LT-HSCs from mice transplanted with 100 Venus⁺ (left) or Venus[−] (right) HSCs. Mean \pm s.d. percentage of Venus⁺ LT-HSCs is shown. **f**, Comparison of transplantation efficiency of single or five HSCs (TCF15-Venus⁺ or Venus[−]). Mean \pm s.d. percentage of myeloid CD45.2⁺ engraftment in recipients ($n = 8$ mice per category) is shown (left). The limiting dilution quantification is also displayed (right). **g**, Model. TCF15 is expressed in a subset of low-output self-renewing HSCs. Upon injury or transplantation, only a subset of these HSCs maintains TCF15 levels, and restores the reservoir pool of relatively quiescent HSCs (some of which can still produce Mk-lineage cells).

TCF15 defines a hierarchy within LT-HSCs

To understand how *Tcf15* expression is regulated in the native context, we generated a knock-in reporter mouse: *Tcf15*-Venus (Extended Data Fig. 9a). Venus fluorescent protein expression was detected in only 0.032% of bone marrow cells and was highly enriched in the LT-HSC compartment, which contained 65.6% of all Lin[−]Venus⁺ cells (Fig. 4a, b, Extended Data Fig. 9b–f). However, consistent with scRNA-seq analysis, *Tcf15* expression within the LT-HSC compartment was markedly heterogeneous, labelling only 38.4% of the cells, and positively correlated with surface receptor levels of endothelial protein C receptor (EPCR; encoded by *Procr*; $r = 0.61 \pm 0.13$) and Sca-1 (encoded by *Ly6a*; $r = 0.65 \pm 0.07$), two markers of quiescent LT-HSCs that were also part of the *Tcf15*^{hi} gene set (Fig. 4c, Extended Data Fig. 9f). To test the functional implications of *Tcf15* expression, we separately transplanted Venus⁺ and Venus[−] LT-HSCs into irradiated recipients (Fig. 4d). Venus⁺ cells reconstituted relatively normal blood and bone marrow compartments, and regenerated both Venus⁺ and Venus[−] HSCs (Fig. 4e, Extended Data Fig. 9g–o). By contrast, Venus[−] cells solely gave rise to Venus[−] cells, displayed relatively impaired primary regeneration and showed significant loss of secondary repopulation capacity (Fig. 4e, Extended Data

Fig. 9g–p). Extreme dilution analysis with single and five-cell transplantation revealed a frequency of approximately one functional HSC for every two cells in the TCF15⁺ LT-HSC compartment, whereas virtually no reconstitution activity was observed in the TCF15⁻ compartment (Fig. 4f). Altogether our analyses indicate that *Tcf15* expression defines a hierarchy within HSCs, where it promotes a self-renewing, quiescent TCF15⁺ cell state with long-term repopulation potential. We propose a model where upon injury or transplantation, a subset of HSCs loses *Tcf15* expression to become active and produce progeny (Fig. 4g).

Recent development of simultaneous lineage and mRNA profiling has enabled direct association of cell behaviours with unique gene expression signatures^{18,56–58}. Applied to haematopoietic regeneration, we have uncovered clone-autonomous stem cell behaviours and the molecular mechanisms that regulate them *in vivo*. We propose that TCF15 is one of the few HSC-restricted transcription factors that specifically regulates the functional LT-HSC state. Our approach may also be directly adapted to study stem cell quiescence regulators in other regenerative tissues.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2503-6>.

- Haas, S., Trumpp, A. & Milsom, M. D. Causes and consequences of hematopoietic stem cell heterogeneity. *Cell Stem Cell* **22**, 627–638 (2018).
- Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**, 418–426 (2018).
- Naik, S. H. et al. Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* **496**, 229–232 (2013).
- Dykstra, B. et al. Long-term propagation of distinct hematopoietic differentiation programs *in vivo*. *Cell Stem Cell* **1**, 218–229 (2007).
- Sun, J. et al. Clonal dynamics of native haematopoiesis. *Nature* **514**, 322–327 (2014).
- Dharampuriya, P. R. et al. Tracking the origin, development, and differentiation of hematopoietic stem cells. *Curr. Opin. Cell Biol.* **49**, 108–115 (2017).
- Kent, D. G. et al. Prospective isolation and molecular characterization of hematopoietic stem cells with durable self-renewal potential. *Blood* **113**, 6342–6350 (2009).
- Carrelha, J. et al. Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells. *Nature* **554**, 106–111 (2018).
- Rodríguez-Fraticelli, A. E. et al. Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212–216 (2018).
- Yamamoto, R. et al. Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. *Cell* **154**, 1112–1126 (2013).
- Yamamoto, R. et al. Large-scale clonal analysis resolves aging of the mouse hematopoietic stem cell compartment. *Cell Stem Cell* **22**, 600–607.e4 (2018).
- Giladi, A. et al. Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat. Cell Biol.* **20**, 836–846 (2018).
- Buenrostro, J. D. et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* **173**, 1535–1548.e16 (2018).
- Dahlin, J. S. et al. A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice. *Blood* **131**, e1–e11 (2018).
- Velten, L. et al. Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19**, 271–281 (2017).
- Cabezas-Wallscheid, N. et al. Vitamin A-retinoic acid signaling regulates hematopoietic stem cell dormancy. *Cell* **169**, 807–823.e19 (2017).
- Wilson, N. K. et al. Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. *Cell Stem Cell* **16**, 712–724 (2015).
- Weinreb, C., Rodriguez-Fraticelli, A. E., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **67**, eaaw3381 (2020).
- Verovskaya, E. et al. Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. *Blood* **122**, 523–532 (2013).
- Cheung, A. M. S. et al. Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice. *Blood* **122**, 3129–3137 (2013).
- Lu, R., Neff, N. F., Quake, S. R. & Weissman, I. L. Tracking single hematopoietic stem cells *in vivo* using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* **29**, 928–933 (2011).
- McKenzie, J. L., Gan, O. I., Doedens, M., Wang, J. C. Y. & Dick, J. E. Individual stem cells with highly variable proliferation and self-renewal properties comprise the human hematopoietic stem cell compartment. *Nat. Immunol.* **7**, 1225–1233 (2006).
- Biasco, L. et al. In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases. *Cell Stem Cell* **19**, 107–119 (2016).
- Scala, S. et al. Dynamics of genetically engineered hematopoietic stem and progenitor cells after autologous transplantation in humans. *Nat. Med.* **24**, 1683–1690 (2018).
- Lu, R., Czechowicz, A., Seita, J., Jiang, D. & Weissman, I. L. Clonal-level lineage commitment pathways of hematopoietic stem cells *in vivo*. *Proc. Natl. Acad. Sci. USA* **116**, 1447–1456 (2019).
- Qian, H. et al. Critical role of thrombopoietin in maintaining adult quiescent hematopoietic stem cells. *Cell Stem Cell* **1**, 671–684 (2007).
- Yoshihara, H. et al. Thrombopoietin/MPL signaling regulates hematopoietic stem cell quiescence and interaction with the osteoblastic niche. *Cell Stem Cell* **1**, 685–697 (2007).
- Kubota, Y., Osawa, M., Jakt, L. M., Yoshikawa, K. & Nishikawa, S. Necdin restricts proliferation of hematopoietic stem cells during hematopoietic regeneration. *Blood* **114**, 4383–4392 (2009).
- Vitali, C. et al. SOCS2 controls proliferation and stemness of hematopoietic cells under stress conditions and its deregulation marks unfavorable acute leukemias. *Cancer Res.* **75**, 2387–2399 (2015).
- Jeong, M. et al. Thioredoxin-interacting protein regulates hematopoietic stem cell quiescence and mobilization under stress conditions. *J. Immunol.* **183**, 2495–2505 (2009).
- Matsumoto, A. et al. p57 is required for quiescence and maintenance of adult hematopoietic stem cells. *Cell Stem Cell* **9**, 262–271 (2011).
- Laurenti, E. et al. Hematopoietic stem cell function and survival depend on c-Myc and N-Myc activity. *Cell Stem Cell* **3**, 611–624 (2008).
- Cabezas-Wallscheid, N. et al. Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell* **15**, 507–522 (2014).
- Lauridsen, F. K. B. et al. Differences in cell cycle status underlie transcriptional heterogeneity in the HSC compartment. *Cell Rep.* **24**, 766–780 (2018).
- Pietras, E. M. et al. Functionally distinct subsets of lineage-biased multipotent progenitors control blood production in normal and regenerative conditions. *Cell Stem Cell* **17**, 35–46 (2015).
- Säwen, P. et al. Murine HSCs contribute actively to native hematopoiesis but with reduced differentiation capacity upon aging. *eLife* **7**, e41258 (2018).
- Yu, V. W. C. et al. Epigenetic memory underlies cell-autonomous heterogeneous behavior of hematopoietic stem cells. *Cell* **168**, 944–945 (2017).
- Balazs, A. B., Fabian, A. J., Esmon, C. T. & Mulligan, R. C. Endothelial protein C receptor (CD201) explicitly identifies hematopoietic stem cells in murine bone marrow. *Blood* **107**, 2317–2321 (2006).
- Pina, C., May, G., Soneji, S., Hong, D. & Enver, T. MLT3 regulates early human erythroid and megakaryocytic cell fate. *Cell Stem Cell* **2**, 264–273 (2008).
- Uckelmann, H. et al. Extracellular matrix protein Matrilin-4 regulates stress-induced HSC proliferation via CXCR4. *J. Exp. Med.* **213**, 1961–1971 (2016).
- Qian, P. et al. Retinoid-sensitive epigenetic regulation of the Hoxb cluster maintains normal hematopoiesis and inhibits leukemogenesis. *Cell Stem Cell* **22**, 740–754.e7 (2018).
- Laurenti, E. et al. CDK6 levels regulate quiescence exit in human hematopoietic stem cells. *Cell Stem Cell* **16**, 302–313 (2015).
- Osawa, M., Hanada, K., Hamada, H. & Nakuchi, H. Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science* **273**, 242–245 (1996).
- Gekas, C. & Graf, T. CD41 expression marks myeloid-biased adult hematopoietic stem cells and increases with age. *Blood* **121**, 4463–4472 (2013).
- Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
- Hill, A. J. et al. On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* **15**, 271–274 (2018).
- Li, W. et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014).
- Rowton, M. et al. Regulation of mesenchymal-to-epithelial transition by PARAXIS during somitogenesis. *Dev. Dyn.* **242**, 1332–1344 (2013).
- Burgess, R., Cserjesi, P., Ligon, K. L. & Olson, E. N. Paraxis: a basic helix-loop-helix protein expressed in paraxial mesoderm and developing somites. *Dev. Biol.* **168**, 296–306 (1995).
- Davies, O. R. et al. Tcf15 primes pluripotent cells for differentiation. *Cell Rep.* **3**, 472–484 (2013).
- Seita, J. et al. Gene Expression Commons: an open platform for absolute gene expression profiling. *PLoS ONE* **7**, e40321 (2012).
- Yamada, T., Park, C. S. & Lacorazza, H. D. Genetic control of quiescence in hematopoietic stem cells. *Cell Cycle* **12**, 2376–2383 (2013).
- Nakamura-Ishizu, A., Takizawa, H. & Suda, T. The analysis, roles and regulation of quiescence in hematopoietic stem cells. *Development* **141**, 4656–4666 (2014).
- Opferman, J. T. et al. Obligate role of anti-apoptotic MCL-1 in the survival of hematopoietic stem cells. *Science* **307**, 1101–1104 (2005).
- Menendez-Gonzalez, J. B. et al. Gata2 as a crucial regulator of stem cells in adult hematopoiesis and acute myeloid leukemia. *Stem Cell Rep.* **13**, 291–306 (2019).
- Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
- Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
- Biddy, B. A. et al. Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564**, 219–224 (2018).

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

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Article

Methods

Animal guidelines

All animal procedures followed relevant guidelines and regulations. All protocols and mouse lines were approved and supervised by the Boston Children's Hospital Institutional Review Board and Institutional Animal Care and Use Committee.

Mice

The TetO-Cas9/M2rtTA mice were a kind gift from S. Orkin (and are available from The Jackson Laboratory, strain no. 029476). To induce Cas9 expression, mice were fed with 1 mg/ml Dox together with 5 mg/ml sucrose in drinking water for the indicated periods of time. Thereafter, Dox was removed. The *Tcf15-Venus* mice were generated from previously described targeted embryonic stem cells⁵⁰. All other mice were the BL/6J strain and obtained from The Jackson Laboratory. Female mice were used as recipients for transplantation. Phlebotomy was performed by retro-orbital sinus peripheral blood collection and analysis (200 µl). Complete blood counts were analysed with an automated Hemacytometer.

Bone marrow preparation

After euthanasia, whole bone marrow (excluding the cranium) of BL6/J or TetO-Cas9/M2rtTA mice was immediately isolated by flushing and crushing in 2% FBS-PBS, and erythrocytes were removed with RBC lysis buffer. CD45.1 (CD45.1, B6.SJL-Ptprca Pep3b/BoyJ, stock no. 002014, The Jackson Laboratory) mice were used as transplantation recipients for CD45.2 (BL6/J) mice.

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 or peripheral blood were purified through four-way sorting using FACS Aria III or FACS Aria Fusion (Becton Dickinson). An example of the sorting strategy for inDrop experiments can be found in Extended Data Fig. 10. Lineage enrichment was performed using anti-cKit (2B8) magnetic beads (Miltenyi Biotec). The following combinations of cell-surface markers were used to define these cell populations: erythroblasts: Ly6G⁻CD19⁻Ter119⁺FSC^{hi}; granulocytes: Ly6G⁺CD19⁻Ter119⁻; monocytes: Ly6C⁺Ly6G⁻CD19⁻Ter119⁻; pro/pre-B cells: Ly6G⁻CD19⁻, Mk progenitors: Lin⁻cKitSca1CD150⁻CD41⁺; LT-HSC: Lin⁻cKitSca1^{+/-}CD150⁻CD48⁻; multipotent progenitors gate 1/short-term HSCs (MP1/ST-HSCs): Lin⁻cKitSca1^{+/-}CD150⁻CD48⁻; MPP2: Lin⁻cKit^{+/-}Sca1^{+/-}CD150⁻CD48⁺; and MPP3/4: Lin⁻cKit^{+/-}Sca1^{+/-}CD150⁻CD48⁺. For cell-cycle analysis, isolated cells were fixed in 4% PFA at room temperature for 10 min and permeabilized with 0.1% Triton X-100 (Sigma) before intracellular staining with 1 µg/ml DAPI and anti-mouse Ki67 antibody. Flow cytometry data were analysed with FlowJo v10 (Tree Star). FACS sorting was performed to obtain the maximal number of available cells from the whole bone marrow extract using purity modes (~98% purity) at ~80% efficiency. Example sorting parameters for LARRY barcoding experiments can be found in Supplementary Fig. 1. The list of antibodies can be found in Supplementary Table 13.

Transplantation assays

LT-HSCs from BL6/J (CD45.2) 8-week-old mice were transplanted in PBS through retro-orbital injection (150 µl per mouse) into CD45.1 recipient mice previously exposed to a lethal gamma radiation dose (2 times 5 Gy with a 2-h interval). Donor cell engraftment (percentage of CD45.2⁺ peripheral blood leukocytes) and labelling frequency were analysed using an LSRII equipment (Becton Dickinson).

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 (700g for 5 min). Sample

DNA was purified by QIAamp DNA Micro kit (56304, Qiagen) and eluted into 10 µl elution buffer before PCR processing. Details for the LARRY pooled library amplification protocol are available at Addgene (no. 140024).

scRNA-seq and low-level data processing

Transcriptome barcoding and preparation of libraries for single-cell mRNA sequencing was performed with inDrop using a 1cellbio device (1cellbio). For our experiment, the eGFP⁺ Lin⁻cKit-enriched bone marrow fraction from recipients was labelled and FACS sorted in four ways to purify SLAM LT-HSCs (Lin⁻Sca1cKit⁺CD150⁺CD48⁻), MPPs (Lin⁻Sca1^{+/-}cKit⁺CD150⁻), MkP (Lin⁻Sca1⁻cKit⁺CD150⁺CD41⁺) and the rest of cKit-enriched cells. All available labelled LT-HSCs are encapsulated in one sample. Then, MPP, MkP and the rest of the cKit-enriched cells are pooled at equal quantities to sample HSC progeny ('KIT' cells). LT-HSC and KIT libraries were processed independently. Libraries for all the populations were prepared the same day, with the same stock of primer-gels and RT-mix, to avoid batch effects. inDrop Primer-gels (v3) were purchased from the Harvard Single Cell Core. Libraries were sequenced on an Illumina NextSeq 500 sequencer using a NextSeq High 75 cycle kit, according to inDrop v3 guidelines (Harvard Single Cell Core). Raw sequencing reads were processed using the inDrop v0.3 pipeline⁵⁹ (github.com/indrops/indrops). LARRY sequencing reads were processed using the LARRY v0.1 pipeline (github/allonkleinlab/LARRY). Single-cell data were analysed and visualized using scanpy v1.4.6⁶⁰ (github/theislab/scanpy) and SPRING v1.6⁶¹ (github/allonkleinlab/SPRING_dev).

Single-cell encapsulation and library preparation for sequencing

For scRNA-seq, we used the inDrop updated protocol previously described⁵⁹, with a modification to allow targeted sequencing of the LARRY barcode. In brief, single cells were encapsulated into 3-nl droplets with hydrogel beads carrying barcoding reverse transcription primers. After reverse transcription in droplets, the emulsion was broken and the bulk material was taken through: (1) second-strand synthesis; (2) linear amplification by in vitro transcription; (3) amplified RNA fragmentation; (4) reverse transcription; and (5) PCR. To specifically amplify barcode-containing eGFP transcripts, we split the amplified RNA fraction (after step (2)) and used one-half for standard library preparation and the other half for targeted lineage barcode enrichment. To target the barcode, we modified the subsequent steps of library prep by (1) skipping RNA fragmentation; (2) priming reverse transcription using a transcript specific primer at 10 mM (TGAGCAAAGACCCCAACGAG); and (3) introducing an extra PCR step using a targeted primer (eight cycles using Kapa HiFi 2X master mix, Roche; primer sequence = TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG NNN Ntaa ccg ttg cta gga gag acc atat) and 1.2× bead purification (Agencourt AMPure XP). Targeted and non-targeted final libraries were pooled at a 1:5 ratio before sequencing.

Read alignment, cell filtering and counts normalization

FASTQ sequence files were demultiplexed and aligned to the GRCm38 mouse reference genome using the inDrops v0.3 pipeline (<https://github.com/indrops/indrops>), generating cell-by-gene counts tables for each experiment and condition. Cells were filtered to include only abundant inDrop barcodes on the basis of visual inspection of the histograms of total transcripts per cell (SPRING_dev/data-prep). The data were further filtered to eliminate putatively stressed or dying cells, defined by having >15% of transcripts coming from mitochondrial genes. We used the SCRUBLET algorithm⁶² (<https://github.com/AllonKleinLab/scrublet>) to inspect putative doublet cells. Cells within each experiment were then normalized (20,000 counts) to have the same total number of transcripts for all subsequent analyses. Filtering and QC parameters (min/max UMIs/cell, median UMIs/cell, normalized

UMIs/cell and median genes/cell) are summarized in Supplementary Table 10.

Generation of SPRING plot layouts

We used SPRING for single-cell data visualization⁶¹. For all SPRING plots shown, we began with total-counts-normalized gene expression data, filtered for highly variable genes using the SPRING gene filter_genes function (from https://github.com/AllonKleinLab/SPRING_dev/blob/master/data_prep/spring_helper.py using parameters (85, 3, 3)), and further filtered to exclude cell-cycle-correlated genes—defined as those with correlation $R > 0.1$ to the gene signature defined by *Ube2c*, *Hmgb2*, *Hmgn2*, *Tuba1b*, *Ccnb1*, *Tubb5*, *Top2a* and *Tubb4b*. To plot cells in SPRING, we embedded cells in 50-dimensional PC space and imported them into SPRING dynamic mode as a k -nearest-neighbour (knn) graph with ($k = 8$). The graph was then allowed to relax in SPRING. To avoid confusion by having many different SPRING plots throughout the manuscript, we reused the single-cell coordinates from two experiments and mapped all other experiments by allowing each cell to choose its 40 nearest neighbours from the first experiment (approximate nearest neighbours), and then take on the average position of the subset of neighbours that were among the original set.

Single-cell clustering

Single-cell transcriptomes were clustered using the Louvain algorithm, following a current recommendation of best practices⁶³. This was performed directly with the SPRING command run_clustering.py, which takes the knn graph and uses the networkx package community.best_partition function to return the most stable partition (resolution was maintained as default = 1). Clusters that were not reproducible between biological replicates were excluded from further analyses. For plotting clusters and populations into single-cell maps, cells were subsampled randomly without substitution (8,000 cells), and plotted top to bottom ordered by clusters (in the following order: '8', '11', '15', '19', '9', '10', '2', '0', '3', '4', '7', '6', '1', '5', '12', '21', '14'). Some clusters were low abundance and not reproducible in independent experiments and these are not shown in these plots ('13', '16', '17', '18', '20', '22', '23').

Cluster annotation

HSC and progenitor clusters were annotated semi-manually, by identifying previously described marker genes among the top cluster-enriched genes (ranked gene z-score test comparing each cluster versus all remaining cells). The full list of cluster markers used are summarized in Supplementary Table 1. HSC clusters were defined by being enriched in the LT-HSC single-cell libraries (compared to the progenitor libraries). Differential gene expression between each HSC subcluster and the rest is shown in Supplementary Table 3. Among the four HSC clusters, HSC-1 presented a gene signature that was closest to the native dormant LT-HSC signature and HSC-2 presented a gene signature that suggested an aged/inflammatory state^{9,12,16,64,65}. By contrast, cluster HSC-3 showed a transcriptional programme associated with HSC cycling and activation^{34,42}, and cluster HSC-4 was defined by markers of activation and Mk priming⁶⁶. From the rest of the cKit⁺ cells, we identified 16 additional clusters containing different progenitor cells, including three stable clusters of MPPs^{67–69}. Progenitor clusters were combined based on the common expression of described lineage markers such as: *Mpo*, *Prtm3* and *Elane* for GM (clusters 1, 2 and 3), *Car1*, *Car2* and *Klf1* for erythroid (Ery-1/2), and *Pf4*, *Itga2b*, *Cd9* and *Rap1b* for Mk progenitors (Mk-1/2). MPP clusters were annotated by being enriched in the progenitor libraries (compared to the LT-HSC libraries) but lacking expression of specific lineage markers as defined.

DPA

For statistical test of the differences in cluster proportions, we used the DPA algorithm⁷⁰. This algorithm returns the probability that an observed distribution of cells among clusters is obtained by random

chance, by shuffling the cells across categories 100,000 times to estimate a null distribution. Clusters with a resulting $P < 0.1$ were considered as significantly differentially enriched between the two conditions.

Cell barcoding with LARRY

The pLARRY vector was constructed by DNA synthesis and Gateway cloning (Vectorbuilder) using a protocol adapted from Naik, Schumacher et al. 2014⁷¹ and Gerrits, Dykstra et al. 2010⁷². The barcoded linker was created by annealing two DNA primers (forward, 5'-CCG CGG ATC CAG ACA TNN NNC TNN NNA CNN NNT CNN NNG TNN NNT GNN NNC ANN ANN ATA TGA GCA ATC CCC ACC CTC CCA CCT AC-3'; reverse, 5'-CTA GGT GGG AGG GTC GGG ATT GCT-3'; IDT DNA). N was a hand mix of 25% A, 25% C, 25% T and 25% G. Primers (10 pmol of each) were mixed in 50 μ l 1× NEB buffer 4 (New England Biolabs). After heating the mixture for 5 min at 95 °C, the primers were allowed to anneal down to 37 °C gradually decreasing the temperature (0.5 °C/min). Then, 1U of Klenow DNA polymerase (3'-5' exonuclease mutant) and 50 nmol of dNTPs were added to the mixture and incubated for 2 h at 37 °C. After Klenow inactivation for 20 min, the barcoded linker was then digested with a mixture of NdeI and BamHI (New England Biolabs) and ligated into the NdeI-BamHI site of the pLARRY vector at a 3:1 ratio. The resulting ligation mix was purified and transformed into 10-beta electroporation ultracompetent *Escherichia coli* cells (New England Biolabs) and grown overnight on LB plates supplemented with 50 μ g/ml ampicillin (Sigma-Aldrich). From eight plates, -0.5–1 × 10⁶ colonies were pooled by flushing plates with LB supplemented with 50 μ g/ml ampicillin. After 6 h of culture, plasmid DNA was extracted with a Maxiprep endotoxin-free kit (Macherey-Nagel). We amplified and sequenced the LARRY library barcodes in bulk (performed in duplicate, with a barcode overlap of 97.7%) and used these sequencing reactions to build a barcode whitelist using the software suite umi-tools (distance = 5). The whitelist is provided in Supplementary Table 12. The pLARRY vector map and plasmid, as well as a sample of the library are available through Addgene (pooled library no. 140024).

LARRY library lentiviral preparation

The LARRY-eGFP library and third-generation lentivirus components (psPAX2 and pMD2.G) were co-transfected into HEK293 Lenti-X cells (Takara Bio) using the TRANS-IT 293 kit (Mirus bio). Lentivirus was harvested every 12 h for 72 h and concentrated using ultracentrifugation (at 50,000g and 4 °C, for 90 min). HEK293 cells were grown in DMEM with 10% FBS and 1% penicillin/streptomycin (GIBCO, Thermo Fisher scientific). HSCs were transduced using spin infection (800g for 90 min at 30 °C) in virus concentrate, cultured at 37 °C for 8 h and then washed out twice with PBS and resuspended in PBS for transplantation.

Calling of lineage barcodes

To call lineage barcodes, we began with an intermediate output of the inDrop pipeline: a list of reads with an annotated cell barcode and unique molecular identifier (UMI). From this list, we extracted all (cell-BC, UMI and lineage-BC) triples that were supported by at least 10 reads, collapsed all lineage-BC's within a hamming distance of four using a graph-connected-components-based algorithm, and carried forward the (cell-BC and lineage-BC) pairs supported by three or more UMIs. To call clones, we then applied a set of filtering steps: (1) cells with the exact same barcode were classified as clones; and (2) pairs of cells in separate sequencing libraries with the same cell-BC and lineage-BC were discarded, since statistically these could only arise from instability of the droplet emulsion. These steps have been implemented in a pipeline available online: <https://github.com/AllonKleinLab/LARRY>. All called barcodes were then verified against the barcode whitelist generated by bulk DNA-seq (see Supplementary Table 12). Typically, we successfully retrieved the lineage barcodes from ~75% to 90% of inDrop GFP⁺ cells using these parameters. Sorting, filtering and barcode retrieval efficiencies are

Article

summarized in Supplementary Table 10. To estimate the quality of our scRNA-seq-based barcode calling approach we verified that: (1) bar-coded and non-barcoded cells present similar transcriptional cluster distributions (variance across clusters was $4.55 \pm 2.78\%$), (2) barcode diversity is sufficient for labelling unique cells, and (3) barcode expression is not significantly silenced even after extended periods of time (Extended Data Fig. 1b–f). We also verified that barcode retrieval efficiency per GFP⁺ cell was similar across populations, with a minor loss of capture efficiency in the preDC and preB clusters (Extended Data Fig. 1g). To further ensure that barcode retrieval by scRNA-seq was representative of the ‘real’ barcode pool, we compared our method with a traditional PCR-based amplification from genomic DNA. We amplified the LARRY barcode from 50 ng of genomic DNA isolated from 200,000 myeloid (Gr1/Mac1⁺) and 100,000 lymphoid (CD19⁺) progenitors, using a nested PCR protocol over three steps with a total of 25 PCR cycles (primers and the PCR protocol are indicated in the following link: <https://benchling.com/s/seq-F1D5aW7t9IBn3q8oywBg>), and sequenced on an Illumina MiSeq. Barcodes were then trimmed, collapsed and compared with the inDrop RNA-seq-derived barcodes (using a hamming distance of four). Analysis revealed that at least ~70% of DNA-seq barcodes (largest barcodes overall) were present in the scRNA-seq data (Extended Data Fig. 1h–j), and the estimated clone sizes derived from scRNA-seq and DNA-seq for each clone were positively correlated ($r = 0.72$; Extended Data Fig. 1k). To further confirm that the low-output activity observed in HSC clones is not due to low barcode sampling efficiency or barcode silencing, we performed a comparison of the relative output calculated from DNA-seq and scRNA-seq data for the same clones, which revealed a significant positive correlation ($r = 0.83$; Extended Data Fig. 1l). We could not robustly retrieve Mk barcodes by DNA-seq, and therefore our estimation of Mk contribution could not be validated in a similar manner. However, our estimations fall in line with previous publications using single HSC transplants, using a more sensitive measurement⁸.

Quantification and classification of HSC clonal behaviours

For each clone, the distribution of cells among clusters was used to quantify three distinct behaviours. Consider N_i the number of all cells, K_i the number of non-HSCs and H_i the number of HSCs, for each clone i (in j clones). For estimating the clone size, we calculated the relative abundance (frequency) of each clone i :

$$f_i = \frac{N_i}{\sum_j N_j}$$

For quantifying the relative output activity (A_i) of each clone i , we divided the frequency in non-HSC clusters (k_i) by the frequency in HSC clusters (h_i). We added a pseudocount of 0.0001 in the denominator to avoid division by 0 in clones without progeny.

$$h_i = \frac{H_i}{\sum_j H_j} k_i = \frac{K_i}{\sum_j K_j} A_i = \frac{k_i}{h_i + 0.0001}$$

For finding statistically significant high-output and low-output clones, we first defined a null hypothesis, assuming no differences in output activity among clones (output = 1). Then, we generated a null hypothesis distribution of A_i values for each clone by sampling 10% of the HSCs (expected progenitors), calculating the A_i for each clone and iterating this process over 1,000 times. We next generated a similar distribution of our observed values, by bootstrapping 10% of the non-HSCs. Finally, for each clone, we compared the two distributions of $A_{i,\text{obs}}$ versus $A_{i,\text{exp}}$ using a two-sample t -test. Clones with $P < 0.05$ were considered as significantly high output or low output and used for subsequent analyses (on average 94.3% of clones).

For quantifying the Mk-lineage bias (B_i) for each clone i , we divided the frequency in Mk clusters ($k_{i,\text{Mk}}$) by the frequency in non-Mk clusters.

We added a pseudocount of 0.0001 in the denominator to avoid division by 0 in clones without progeny.

$$B_{i,\text{Mk}} = \frac{k_{i,\text{Mk}}}{k_{i,\text{non-Mk}} + 0.0001}$$

For finding statistically significant Mk-biased clones, we first defined a null hypothesis, assuming no differences in Mk bias among clones ($B_i = 1$). Then, we generated a null hypothesis distribution of B_i values for each clone by sampling 10% non-Mk progenitors (expected Mk), calculating the B_i for each clone and iterating this process over 1,000 times. We next generated a similar distribution of our observed values, bootstrapping 10% of the Mk progenitors. Finally, for each clone, we compared the distributions of $B_{i,\text{obs}}$ and $B_{i,\text{exp}}$ using a two-sample t -test. Clones with $P < 0.05$ and $B_i > 1$ or $B_i < 1$ were considered as significantly biased and used for subsequent analyses. For calculating signatures, we considered $B_i > 1$, but quantification of clones with $B_i < 1$ are also shown in Fig. 1j and Supplementary Table 2. For plotting these measurements into single-cell maps, cells were subsampled randomly without substitution (8,000 cells), and then ordered top to bottom, first by clusters (1–23) and then randomly within each cluster. For the separate plots of high-output and low-output clones in Extended Data Fig. 2d, e, cells were subsampled randomly without substitution (2,200 cells) and then ordered (top to bottom) in the same way. The results of all these quantifications are summarized in Supplementary Table 11.

Single-cell differential gene expression analysis

Single-cell differential gene expression was carried out with scanpy, using the rank_genes_groups function, which performs a t -test with Benjamini–Hochberg correction for multiple testing. The numbers of cells used for each comparison are summarized in each corresponding Supplementary Table. Symmetrically opposite gene expression analysis of sgTcf15 and TetO-Tcf15 HSCs was performed by multiplying the scores of each differentially expressed gene ($\text{rank}_{\text{sgTcf15}} \times \text{rank}_{\text{TetO-Tcf15}}$), selecting all results with a negative sign (those expressed in opposite directions) and then further filtering those downregulated in sgTcf15, with the assumption that these genes are regulated positively by TCF15 transcription factor activity. The resulting list was analysed using Topogene⁷³ for gene ontology analysis and is shown in Supplementary Table 9.

Gene signature scores

Scores for gene signatures were generated with the scanpy score_genes function, with default options. Selected genes to build each score were the top differentially enriched genes (adjusted $P < 0.05$) after ranking by combined score. These genes are indicated in Supplementary Table 2. A similar approach was taken for computing previously published stem cell signatures. The Wilson et al. MolO and suMo signatures¹⁷ and the Giladi et al. StemScore signatures¹² were used as described in their respective publications. For the Pietras et al. HSC signature³⁵, we used the top 1,000 genes with adjusted $P < 0.05$. For the Cabezas-Wallscheid et al. 2014 (HSC)³³ and 2017 (dHSC)¹⁶ signatures, we used all of the genes with adjusted $P < 0.05$ (273 and 787 genes, respectively). For the Lauridsen et al. RA-CFPdim HSCs³⁴, we used the genes with adjusted $P < 0.05$. Signature gene lists from these publications are shown in Supplementary Table 2. For plotting these signature scores into single-cell maps, cells were plotted ordered by signature score (highest score on top).

Secondary transplantation of barcoded HSCs

eGFP⁺ immunophenotypic HSCs from barcoded primary transplants (~7,500 cells) were isolated by FACS in 2% FBS-supplemented PBS and split randomly at equal proportions into two microcentrifuge tubes. Cells from one tube were prepared and analysed using inDrop as previously indicated. The HSCs from the remaining tube were spun down, resuspended in 300 μ l PBS and injected retro-orbitally into two lethally

irradiated CD45.1 BL6 mice (secondary recipients). Secondary recipients were analysed (100 µl retro-orbital blood) to verify engraftment after 2 and 4 months. After 4 months, secondary recipients were euthanized and all LT-HSCs were purified similar to the primary transplants and analysed by inDrop independently. For each recipient, a fraction of the cKit⁺ progenitors was also analysed by inDrop to determine the contribution of clones to differentiated blood lineages. The pipeline to analyse secondary transplant data is available at <https://github.com/AllonKleinLab/StemCellTransplantationModel> and a more extensive description of mathematical methods and results can be found in Supplementary Methods.

CROP-seq CRISPR screening

To select the candidate genes, we ranked all genes expressed by low-output HSC clones, excluded those that were not specific to the LT-HSC compartment, and then further excluded most genes previously described to have a role in HSC maintenance, to focus on novel discoveries (Supplementary Table 4). This selection allowed us to focus on discovering new candidates of steady-state stem cell quiescence. We included two genes, which have been described to have an HSC activation (loss of quiescence) phenotype upon knockout (*Ptger4* and *Tsc22d1*), as putative positive controls. The final sgRNA library (carrying three sgRNAs per each candidate, and five control sgRNAs; Supplementary Table 5) was cloned into a custom-made CROP-seq-mNeonGreen vector using the published protocol in <http://crop-seq.computational-epigenetics.org>.

We isolated *TetO-Cas9;M2rtTA* LT-HSCs and transduced them with the library (MOI = 0.3) for 8 h. We then transplanted the cells into six separate recipients (in two independent experiments), waited until steady-state reconstitution (16 weeks) and added Dox in drinking water for up to 2 months. We analysed the blood of recipients before and after Dox addition, and sort-purified different bone marrow populations for deep sgRNA sequencing at the end point. Finally, we used inDrop to encapsulate all of the available LT-HSCs (18,630 cells) and a fraction of the remaining cKit⁺ cells (22,426 cells) to sample different progenitors. To sequence the sgRNAs, we followed the published protocol in Datlinger et al.⁴⁵, adapting it to inDrop sequencing primers by adding the inDrop adapters for inDrop multiplexing and mixing, as performed for the LARRY barcode, and modifying the LARRY barcode calling pipeline. Crop-seq sgRNA bulk sequencing from DNA was also performed as indicated in Datlinger et al.⁴⁵, using up to 10 ng of DNA purified from sorted immunophenotypic gates, or 10 ng of lentiviral plasmid library maxiprep. Libraries were indexed using TruSeq Illumina primers and sequenced on Illumina NextSeq 500. Sequences were demultiplexed and aligned to a custom bowtie index containing the sgRNA sequences for the whole library. Reads were then mapped using bowtie, sorted, counted and normalized to 1,000,000 counts per index. Bulk sgRNA sequence enrichment was performed using MAGeCK⁴⁷.

Statistical methods

Statistical analysis tests, parameters and results are described in each corresponding figure, with details in specific sections of the Methods, as indicated. The description of statistical and mathematical methods for data analysis of secondary transplants is included in the Supplementary Information (Supplementary Methods).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Raw data and counts matrices are available at the GEO (GSE134242). The LARRY barcoding tool is available at Addgene (no. 140024).

Data analyses are available at the following links: http://github.com/rodriguez-fraticelli/Tcf15_HSCs and <https://github.com/AllonKleinLab/StemCellTransplantationModel>. Source data are provided with this paper.

Code availability

Code, processed data and analyses are available at http://github.com/rodriguez-fraticelli/Tcf15_HSCs and <https://github.com/AllonKleinLab/StemCellTransplantationModel>.

59. Zilionis, R. et al. Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* **12**, 44–73 (2017).
60. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
61. Weinreb, C., Wolock, S. & Klein, A. M. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* **34**, 1246–1248 (2018).
62. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291.e9 (2019).
63. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
64. Kirschner, K. et al. Proliferation drives aging-related functional decline in a subpopulation of the hematopoietic stem cell compartment. *Cell Rep.* **19**, 1503–1511 (2017).
65. Wilson, A. et al. Hematopoietic stem cells reversibly switch from dormancy to renewal during homeostasis and repair. *Cell* **135**, 1118–1129 (2008).
66. Haas, S. et al. Inflammation-induced emergency megakaryopoiesis driven by hematopoietic stem cell-like megakaryocyte progenitors. *Cell Stem Cell* **17**, 422–434 (2015).
67. Tusi, B. K. et al. Population snapshots predict early hematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018).
68. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
69. Grimes, H. L. et al. Single cell transcriptome-based dissection of lineage fate decisions in myelopoiesis. *Exp. Hematol.* **42**, S21 (2014).
70. Farbehi, N. et al. Single-cell expression profiling reveals dynamic flux of cardiac stromal, vascular and immune cells in health and injury. *eLife* **8**, e43882 (2019).
71. Naik, S. H., Schumacher, T. N. & Périé, L. Cellular barcoding: a technical appraisal. *Exp. Hematol.* **42**, 598–608 (2014).
72. Gerrits, A. et al. Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood* **115**, 2610–2618 (2010).
73. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–W311 (2009).

Acknowledgements A.E.R.-F. acknowledges support by the Life Sciences Research Foundation Merck Fellowship, the European Molecular Biology Organization Long-term Fellowship (ALTF 675-2015), and the NIH NHLBI K99/R00 award (K99HL146983). A.E.R.-F is a Scholar of the American Society of Hematology and a Special Fellow of the Leukemia and Lymphoma Society (3391-19). S.-WW. is supported by a Damon Runyon Cancer Research Foundation Computational Biology Fellowship. A.M.K., S.-WW. and C.W. acknowledge support by NIH grants R33CA212697-01 and 1R01HL14102-01, the Harvard Stem Cell Institute Blood Program Pilot grant DP-0174-18-00, and the Chan-Zuckerberg Initiative grant 2018-182714. S.L. was supported by a Senior Fellowship from the Wellcome Trust WT103789AIA. F.D.C. was supported by NIH grants HL128850-01A1 and P01HL13147. F.D.C. is a scholar of the Howard Hughes Medical Institute and the Leukemia and Lymphoma Society. We acknowledge C.-Y. Lin for generating *Tcf15-Venus* mice; the assistance of R. Mathieu and the Flow Cytometry Core at Boston Children's Hospital; the assistance of A. Ratner and the Harvard Medical School Single Cell Core; the assistance of the Harvard Biopolymers Facility for high-throughput sequencing; and members of the Camargo and Klein lab for helpful discussions, method development and scripts. Illustrations were created with BioRender.

Author contributions A.E.R.-F. performed library generation, lentiviral preparation, cell barcoding, flow cytometry and sorting, stem cell transplants, bleeding, bone marrow preparation, scRNA-seq, library preparation and single-cell data analysis. M.J. performed lentiviral preparation and cell barcoding. C.W. performed single-cell library preparation, script development and single-cell data analysis. S.-WW. and A.M.K. generated the null-equivalent model and analysed the secondary transplantation data. A.E.R.-F. and M.U. carried out the Crop-seq experiments. S.L. and R.P.M. generated the *Tcf15-Venus* knock-in mouse and isolated and provided bone marrow. A.E.R.-F., C.W., S.-WW., A.M.K. and F.D.C. contributed to writing the manuscript. A.M.K. supervised S.-WW. and C.W. F.D.C. and A.E.R.-F. supervised the study.

Competing interests A.M.K. is a co-founder of 1cellbio, Ltd. The other authors declare no competing interests.

Additional information

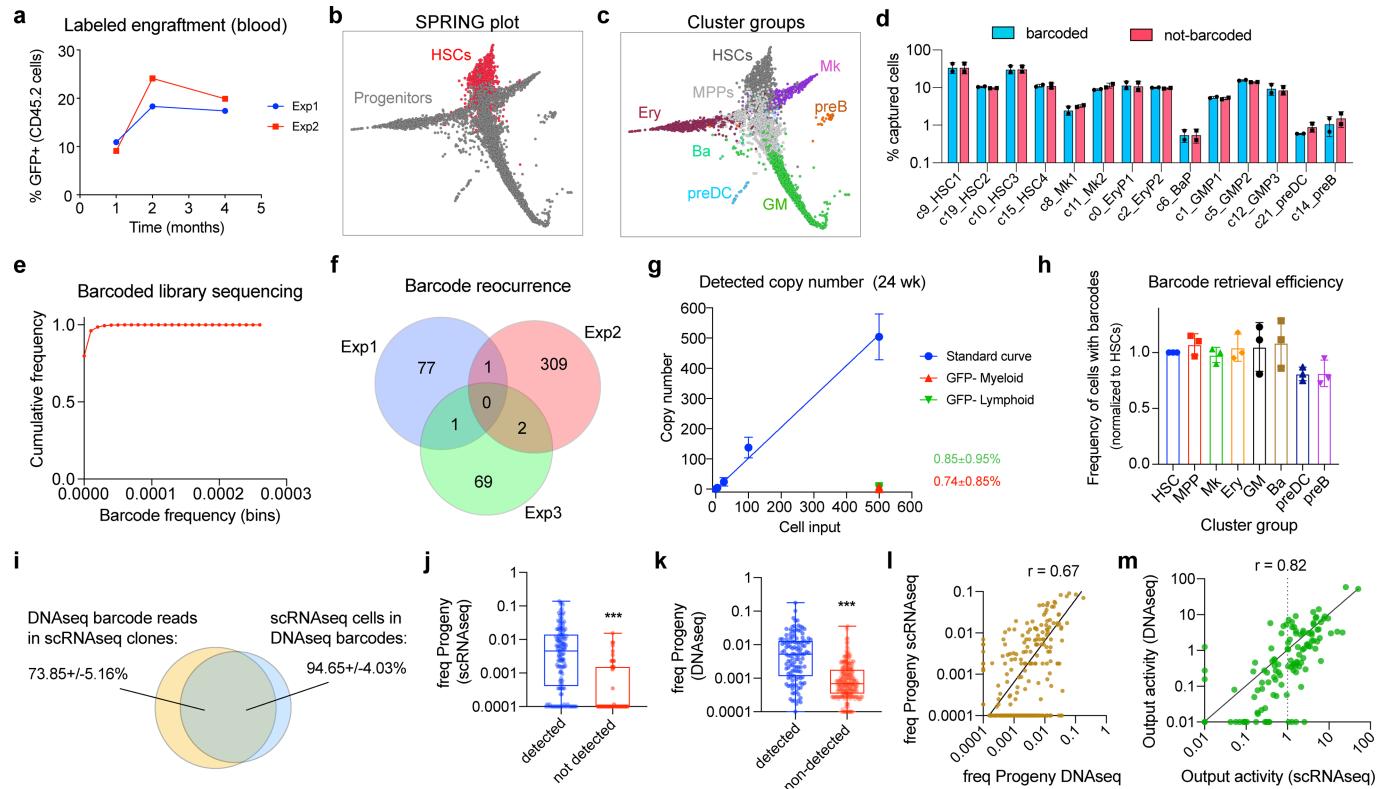
Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2503-6>.

Correspondence and requests for materials should be addressed to F.D.C.

Peer review information *Nature* thanks Thomas Höfer, Samantha A. Morris and Leila Périé for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

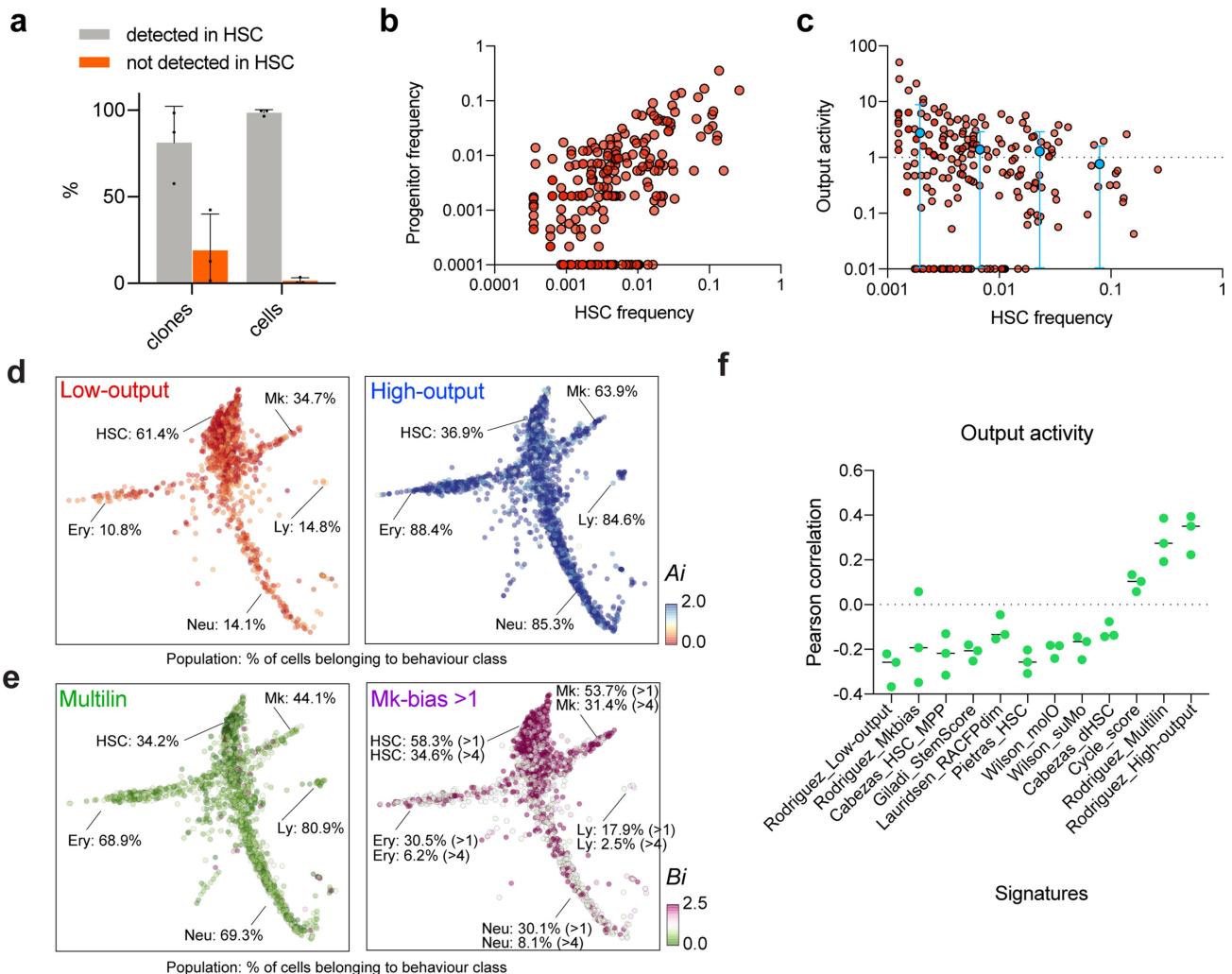
Article



Extended Data Fig. 1 | Controls and validation of the approach.

a, Comparison of peripheral blood engraftment for barcode-expressing cells (EGFP+) in two representative experiments. **b**, Merged cluster labelling of the data set, indicating the localization of HSCs (pink) and Progenitors (grey) in the single cell map plotted using SPRING. **c**, Merged cluster labelling, indicating the localization of Erythroid (Ery), Basophil (Ba), Dendritic cell (preDC), Granulocyte-Monocyte (GM), B-cell (preB) and Megakaryocyte (Mk) progenitors. **d**, Cluster distribution comparison of barcoded (blue) and non-barcoded (red) cells. Mean \pm s.d. % of cells assigned to each cluster ($n = 2$ independent experiments). **e**, Barcode library diversity estimation, showing cumulative barcode frequency at different barcode abundances (binned). 96% of the library is represented by barcodes with a freq < 0.00001. **f**, Barcode library diversity estimation, showing the barcode overlap between independent experiments. Average overlap is 1.3%. **g**, Barcode silencing estimation, showing the % of barcodes detected in the genomic DNA of EGFP-negative cells by quantitative PCR. A calibration curve using sorted numbers of EGFP-positive cells is shown in blue. Mean \pm s.d. of $n = 3$ independent animals are shown. Lines represent linear regression from the data. **h**, Differences in barcode detection efficiency. The histogram represents the proportion of barcoded cells in each population as detected by scRNaseq

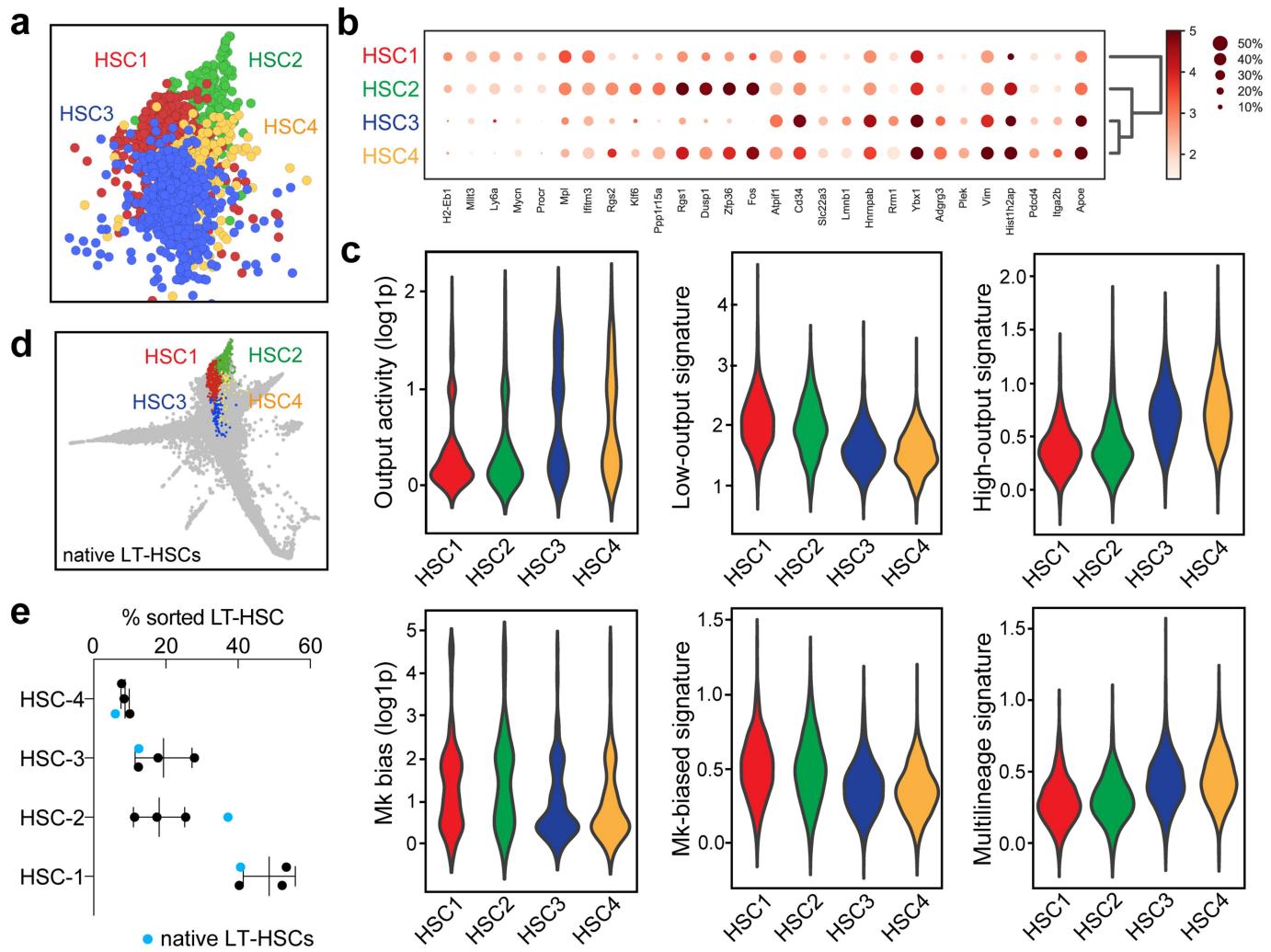
(HSCs, MPP, Mk, GM, Ery, Ba, preDC and preB). Data shown are mean \pm s.d. from 3 independent experiments. The data are shown normalized by the proportion of barcoded HSCs ($72.3\% \pm 5.5\%$). The mean efficiency drops for the preDC and preB populations, but it is not significant (paired two-sided t -tests, $P = 0.07$, $P = 0.17$). **i**, Mean \pm s.d. % of shared DNAseq reads and scRNaseq cells across barcodes in progenitors ($n = 3$ independent experiments). **j**, Distribution of progeny frequencies for all clones (quantified by scRNaseq), and labelled according to their presence or absence in DNAseq barcodes. Box plot shows median and interquartile range. Error bars are min/max values. *** $P < 0.01$ two-sided t -test ($n_{detected} = 137$, $n_{not_detected} = 50$). **k**, Distribution of progeny frequencies for all barcodes (quantified by DNAseq), and labelled according to their presence or absence in scRNaseq-recovered barcodes. Box plot shows median and interquartile range. Error bars are min/max values. *** $P < 0.01$ two-sided t -test ($n_{detected} = 127$, $n_{not_detected} = 286$). **l**, Correlation of DNAseq and RNA-seq barcode frequencies ($n = 429$). Pearson correlation (r) is shown. Line represents simple linear regression of the data. A pseudocount of 0.0001 is used for plotting clones undetected in either set. **m**, Correlation of DNAseq and RNA-seq measurements of HSC output activity for all HSC clones ($n = 136$). Pearson correlation (r) is shown. Line represents simple linear regression of the data. A pseudocount of 0.01 is used to plot clones with output = 0.



Extended Data Fig. 2 | Description of HSC heterogeneity according to their output activity and clone size. **a**, Histogram showing % of cells (right) and % of clones (left) in progenitors that are not detected in HSCs ($n=3$ independent experiments). Whereas some clones are not detected in HSCs (orange bar, left), these are typically single cell clones and minimally contribute to progenitor cellularity (orange bar, right). $p_{\text{clones}} = 0.022$ and $p_{\text{cells}} < 0.001$. Holm-Sidak multiple-test corrected t -test. **b**, Scatter plot showing correlation between HSC clone size, h_i (expressed as fraction of total HSCs in each experiment), and clonal output activity, k_i (fraction of total progenitors), for each detected clone (data are pooled from 5 mice). Pearson correlation $r = 0.59$ ($n = 226$ clones, from all 3 independent experiments). A pseudocount of 0.0001 is used for progeny frequency to display the zeros (clones with no output). **c**, Scatter plot showing HSC clone sizes and their range of differentiated output activity. Pearson correlation $r = -0.097$ (slope non-significantly different than zero, $P = 0.1449$, $n = 226$ clones). A pseudocount of 0.01 is used for output activity to display clones for which progeny is not detected. The binned average and range are

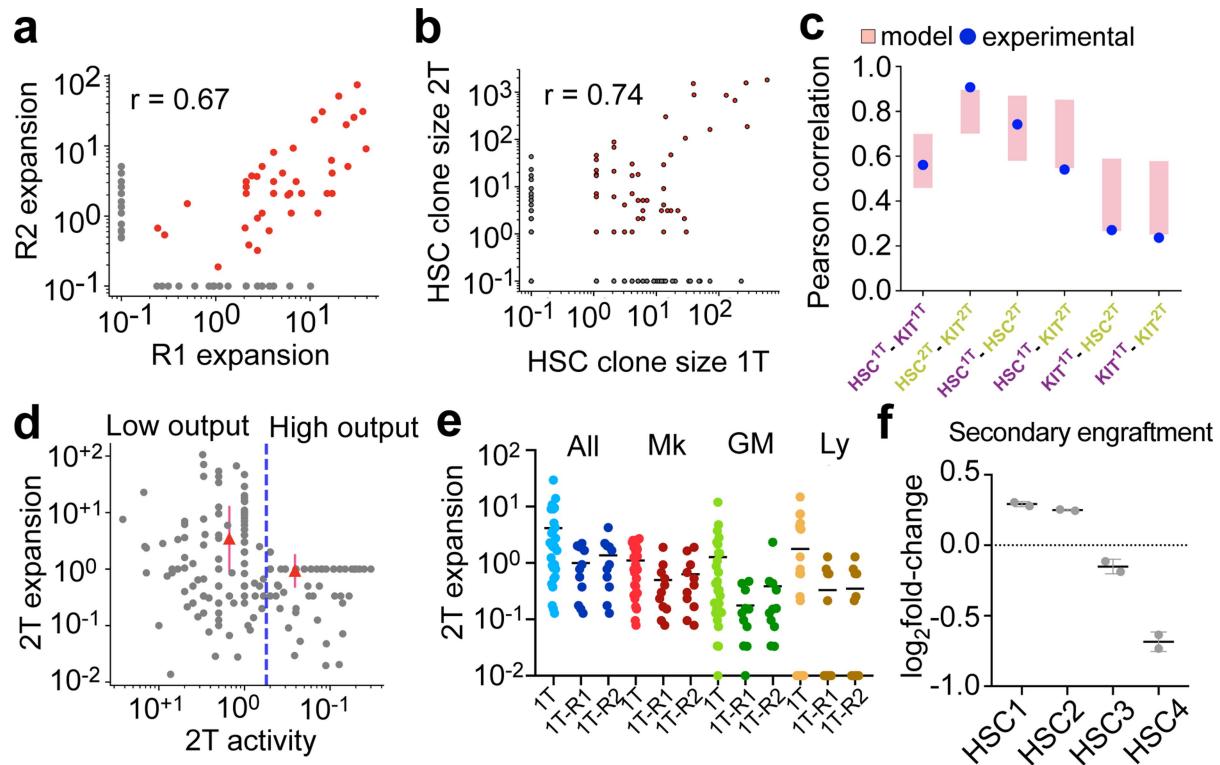
shown in blue (HSC frequency bins are $[0.0001-0.005], n=127, [0.005-0.01], n=33, [0.01-0.05], n=52 [0.05-1], n=14$). **d**, Single cell maps showing the clonal HSC output activity values for each single cell. Low-output clones are shown on the left and high-output clones are shown on the right. For each population (HSCs, Mk, Ery, Ly and Neu), the percentage of cells that belongs to clones of the indicated behaviour class is shown. Scale range, 0 (red) to 2 or more (blue). Plotted single cells are randomly subsampled ($n = 2000$) without replacement. **e**, Single cell maps showing the clonal HSC Mk-bias values for each single cell. Non-biased multilineage clones are shown on the left and Mk-biased (bias >1) clones are shown on the right. For each population (HSCs, Mk, Ery, Ly and Neu), the percentage of cells that belongs to clones of the indicated behaviour class is shown. Scale range, 0 (green) to 2.5 or more (pink). Plotted single cells are randomly subsampled ($n = 2000$) without replacement. **f**, Pearson correlation between the output activity and the average signature score of each clone, for different computed signatures as in Fig. 1. Black bars indicate mean of 3 independent experiments.

Article



Extended Data Fig. 3 | Description of HSC subclusters. **a**, SPRING plot showing the localization of the four reproducible HSC subclusters, HSC1-4. The plot is representative of one of three experiments with similar results. **b**, Marker gene expression for HSC subclusters. **c**, Violin plots showing the values for output activity, Mk-bias, and the scores of different HSC behaviour signatures. Violin plots show all the data (min-to-max) and are representative from one of 3 independent experiments ($n_{HSC1} = 2206$, $n_{HSC2} = 577$, $n_{HSC3} = 1794$, $n_{HSC4} = 649$). DPA results (P -values) are indicated for each HSC cluster in order

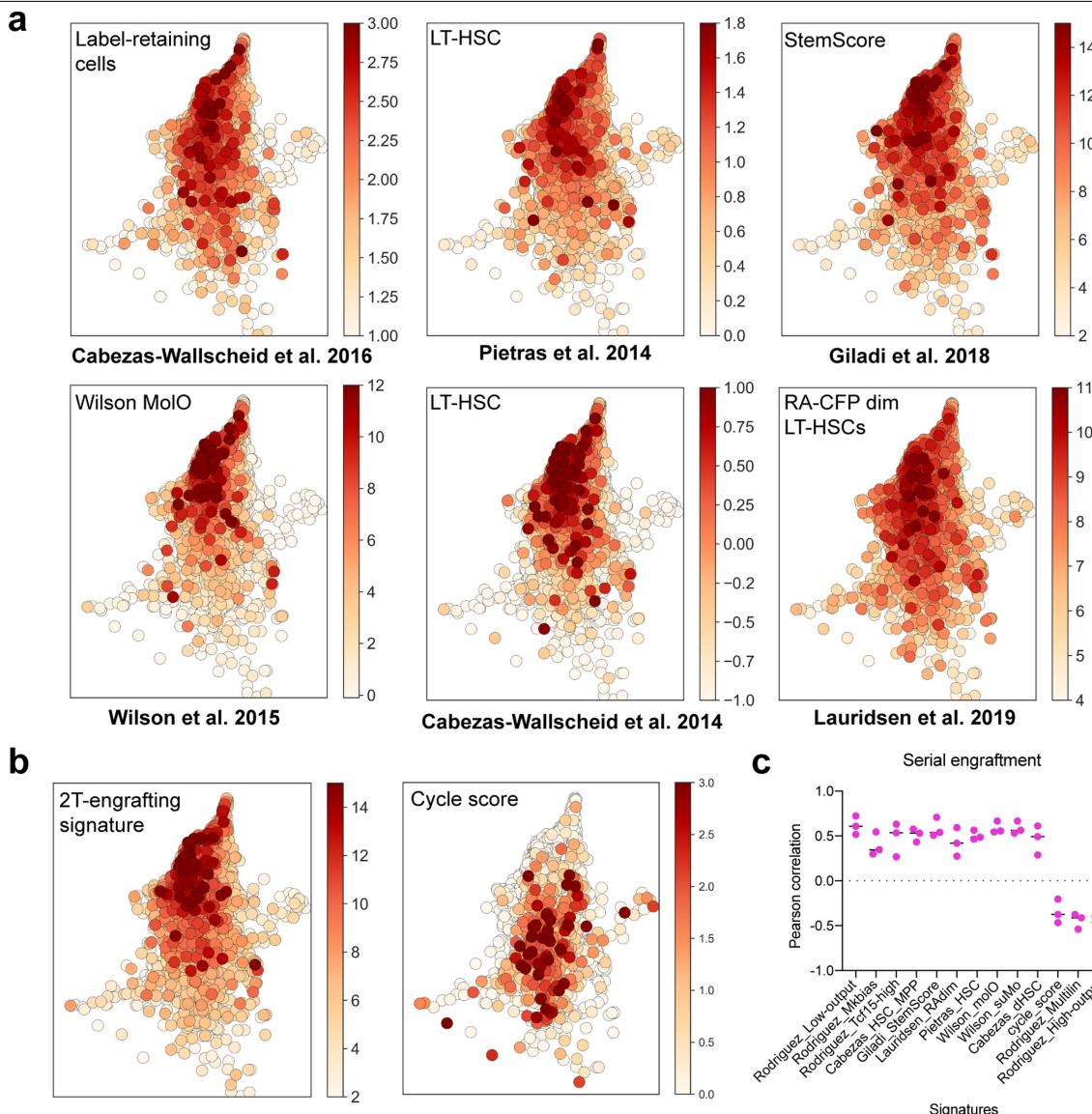
from HSC1 to HSC4. Low-output: 0.0023, 0.0051, <0.0001 , 0.0114. High-output: <0.0001 , 0.3883, <0.0001 , 0.0006. Mk-bias: 0.0002, 0.0172, 0.0516, 0.0182. Multilineage: 0.2257, 0.0763, 0.4374, 0.1977. **d**, SPRING plot showing distribution of native LT-HSCs ($n=1$) mapped by approximate nearest neighbours (see Methods). **e**, Cluster distribution of native LT-HSCs (blue dots) compared to transplant HSCs (black dots). Mean \pm S.D., $n=3$. Chi-square test (transplant HSCs vs. native LT-HSCs), $P_{exp1} = 10^{-8}$, $P_{exp2} = 0.0007$, $P_{exp3} = 0.0483$.



Extended Data Fig. 4 | Additional data for validation of the null-equipotent HSC model.

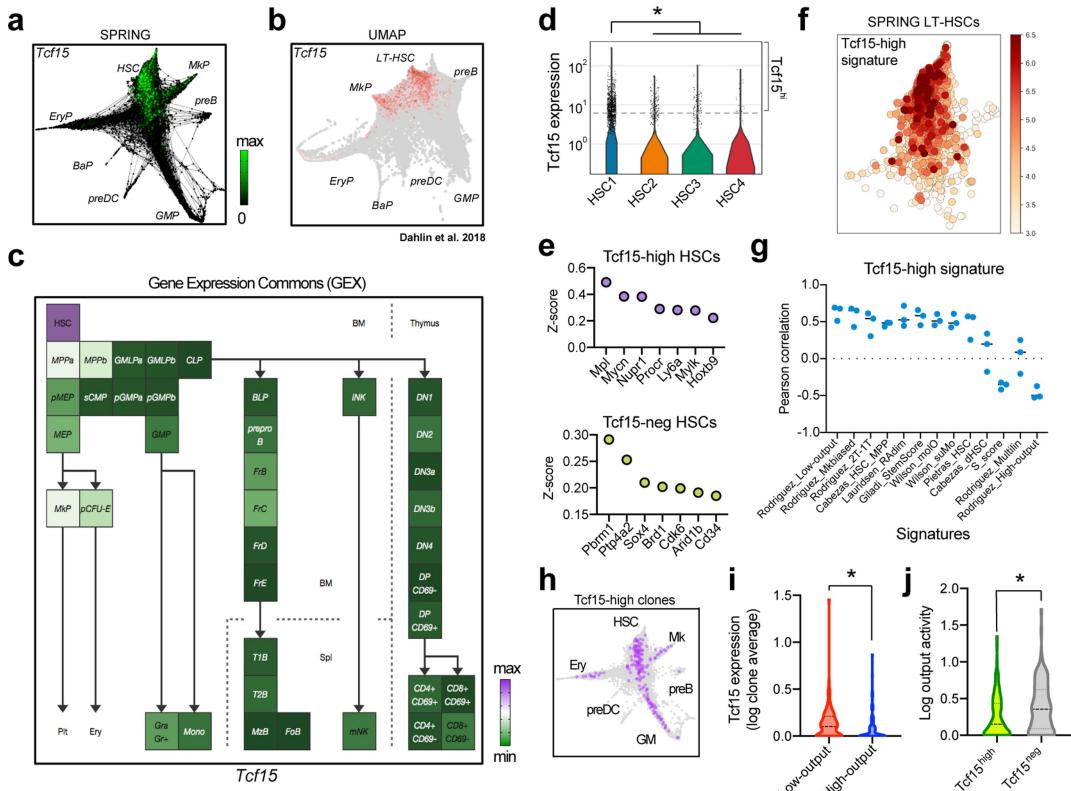
a, Scatter plot showing the Pearson correlation between expansion of HSC clones in each secondary recipient (R1 and R2, $n=133$ clones). **b**, Scatter plot showing the Pearson correlation between HSC clone size in primary and secondary recipients ($n=485$ clones). The grey dots are clones only detected in either primary or secondary recipients, using pseudocount of 0.1 to plot in logarithmic scale. **c**, Histogram depicting the values for clone size correlations between the designated populations. The experimental data are shown in blue, and the data (range) from the null equipotent model is shown in pink (1σ). **d**, Scatter plot of relative HSC output activity in the primary transplant (1T output) vs. clone expansion in secondary recipients (2T expansion). Clonal expansion (2T/1T clone size) is used, instead of absolute clone size, to account for the effect of 1T clone size on the estimation of

engraftment capacity. To avoid numerical divergence, pseudocount = 1 is added before taking the ratio. High-output clones are top 40% clones ranked by their 1T activities, and the remaining 60% are classified as low-output clones. Red triangles show the mean \pm s.d. 2T expansion for each category ($n=485$ clones, combined from both recipients). **e**, Scatter plot showing relative 1T output activity across different lineages for all 1T clones and secondary engrafting clones (R1 and R2 shown separately). Bar indicates mean output value. **f**, Fold-change in the HSC cluster distribution showing the enrichment of secondary transplantation capacity in HSC-1/2/3/4 subclusters. Bars indicate mean \pm s.d. ($n=2$). Chi-square test $P=0.009$ (observed vs. expected distribution). See data availability statement for source data of secondary transplantation assays.



Extended Data Fig. 5 | Comparison of LT-HSC signatures. **a**, Single cell plots of transplanted and barcoded HSCs showing the scores of previously published HSC signatures. Pietras et al. 2014 HSC signature is derived from comparison of Flt3-CD48-CD150+ LSKs (HSCs) versus all other progenitor populations. Lauridsen et al. 2019 dormant HSC (dHSC) signature is derived from comparison of RA-CFPdim HSCs, which are enriched in quiescent HSCs, versus RA-CFPpositive HSCs, which are enriched in cycling HSCs. Giladi et al. 2018 StemScore is derived from single cell data analysis of genes correlating with *Hif* expression in naive HSCs. Wilson et al. 2015 MolO signature is derived from single cell expression data of index-sorted LT-HSCs. Cabezas-Wallscheid et al.

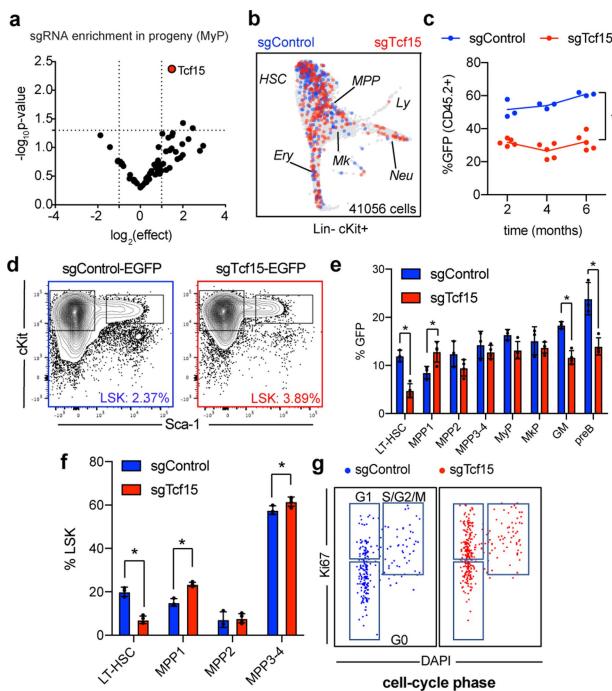
2017 label-retaining HSC signature is derived all HSC genes significantly upregulated in H2B-GFP^{hi} label-retaining HSCs, compared to H2B-GFP^{low}. **b**, Single cell plot showing the 2T-engrafting signature score, derived from the comparison of serially repopulating HSC clones and non-serially repopulating clones (Fig. 2). **c**, Pearson correlation between the 2T-engraftment long-term repopulating signature score and the indicated HSC signature scores. Low-output, high-output, Mk-biased and Multilineage signature scores are derived from the analyses shown in Fig. 1. Black bars indicate mean of 3 independent experiments.



Extended Data Fig. 6 | *Tcf15* expression is restricted to HSCs, and it is highest in the low-output clones. **a**, Localization of expression of *Tcf15* along the single cell manifold using SPRING. Major cluster groups are labelled. The plot shows cells from one of 3 experiments with similar results ($n = 16976$ cells). **b**, Localization of expression of *Tcf15* along the single cell manifold in the Dahlin et al. 2018 data set using Scanpy ($n = 44802$ cells pooled from 6 animals). Major cluster groups are labelled. **c**, Localization of *Tcf15* expression along the bone marrow FACS-pure populations in Gene Expression Commons. **d**, Expression levels of *Tcf15* in the different HSC subclusters. Violin plots show all the data (min-to-max). The scale (width) of the violin plot is adjusted to show the same total area for each subcluster ($n_{HSC1} = 10815$, $n_{HSC2} = 2265$, $n_{HSC3} = 2867$, $n_{HSC4} = 900$). *Tcf15* expression scale is log (normalized UMI). DPA results (P -values) testing enrichment of *Tcf15*^{hi} (>5 UMI) cells across each HSC cluster are, in order, from cluster HSC1 to HSC4: <0.0001, 0.4843, <0.0001, 0.0009. *indicates enrichment in HSC1. **e**, Selected genes enriched in *Tcf15*^{hi} HSCs and *Tcf15*^{neg} HSCs. **f**, Single cell plot of the *Tcf15*^{hi} signature score, using genes enriched in

Tcf15-expressing cells (z -score >0.3). **g**, Pearson correlation between the *Tcf15*^{hi} signature score and the indicated HSC signature scores. Bars indicate average of $n = 3$ independent experiments. Low-output, high-output, Mk-biased and Multilineage signature scores are derived from the analyses shown in Fig. 1. **h**, SPRING plots showing distribution of *Tcf15*^{hi} HSC clones and their progeny (purple) compared to the rest of HSCs (light grey) in primary transplants. Major cluster groups are labelled. Cells shown are from a representative experiment of 3 independent experiments with similar results ($n = 16976$ cells). **i**, Violin plot showing the average distribution of *Tcf15* expression levels in low-output ($n = 123$) versus high-output ($n = 101$) HSC clones taken from 3 independent experiments with similar results. Violin plot shows all data, with median (dashed line) and quartiles (dotted lines). * $P = 0.0165$ (two-sided unpaired t -test). **j**, Violin plot showing the distribution of relative output activity in *Tcf15*^{hi} ($n = 95$) versus *Tcf15*^{neg} ($n = 129$) HSC clones. Violin plot shows all data, with median (dashed line) and quartiles (dotted lines). * $P = 0.0015$ (two-sided unpaired t -test).

Article



Extended Data Fig. 7 | Additional measurements on Tcf15 requirement for HSC quiescence. **a**, Volcano-plot showing the multiple comparison-corrected (Bonferroni) unique *t*-test for each gene in a representative population (LSK⁺CD41⁺; Myeloid progenitors). Two-sided test, $n = 6$ independent mice.

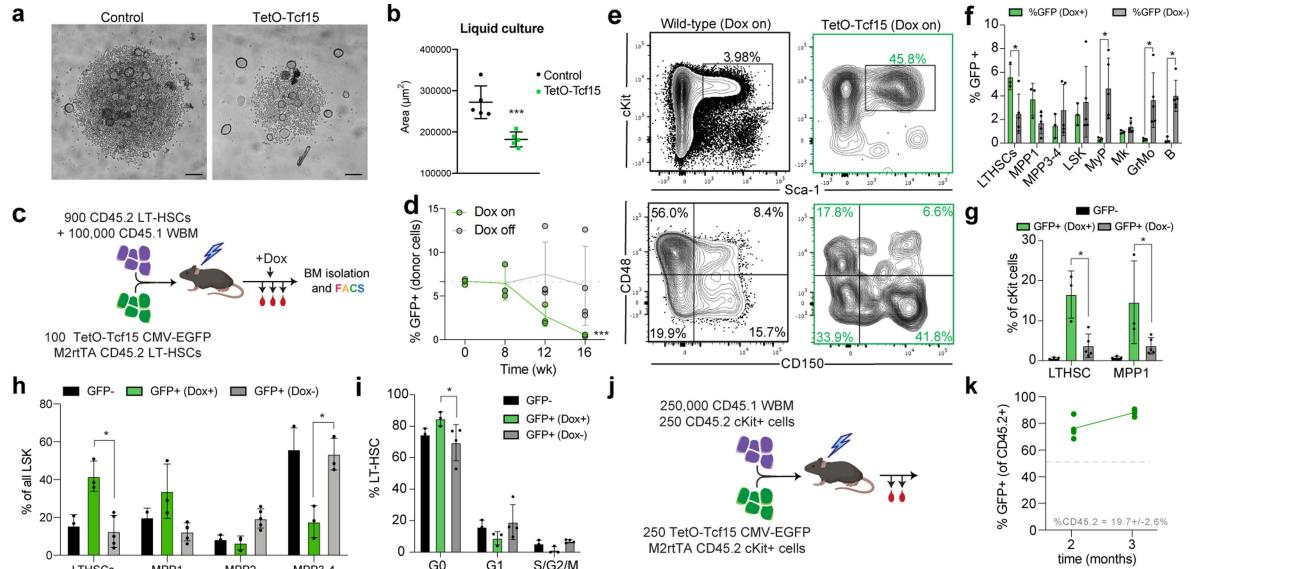
b, SPRING plot localization of sgControl vs. sgTcf15 cells using inDrop. Identified branches are labelled by marker gene expression. Plot is representative from one of $n = 2$ independent single-cell experiments (each experiment from 3 mice combined). **c**, Quantification of peripheral blood engraftment as %EGFP+ cells (of all CD45.2+), comparing sgControl (blue) and sgTcf15 (red) donor cells.

* $P = 0.0017$ (two-sided unpaired *t*-test, $n_{\text{sgControl}} = 4$ and $n_{\text{sgTcf15}} = 5$ animals). Lines indicate mean per group. **d**, FACS plots showing Lin- cKit-enriched bone marrow staining for LSKs in primary recipients. Only EGFP+ cells are shown in the plots. Plots are taken from representative one animal per group from $n = 3$ experiments. **e**, Quantification of bone-marrow engraftment as Mean \pm s.d.

%EGFP+ cells (of all bone marrow) in each designated compartment.

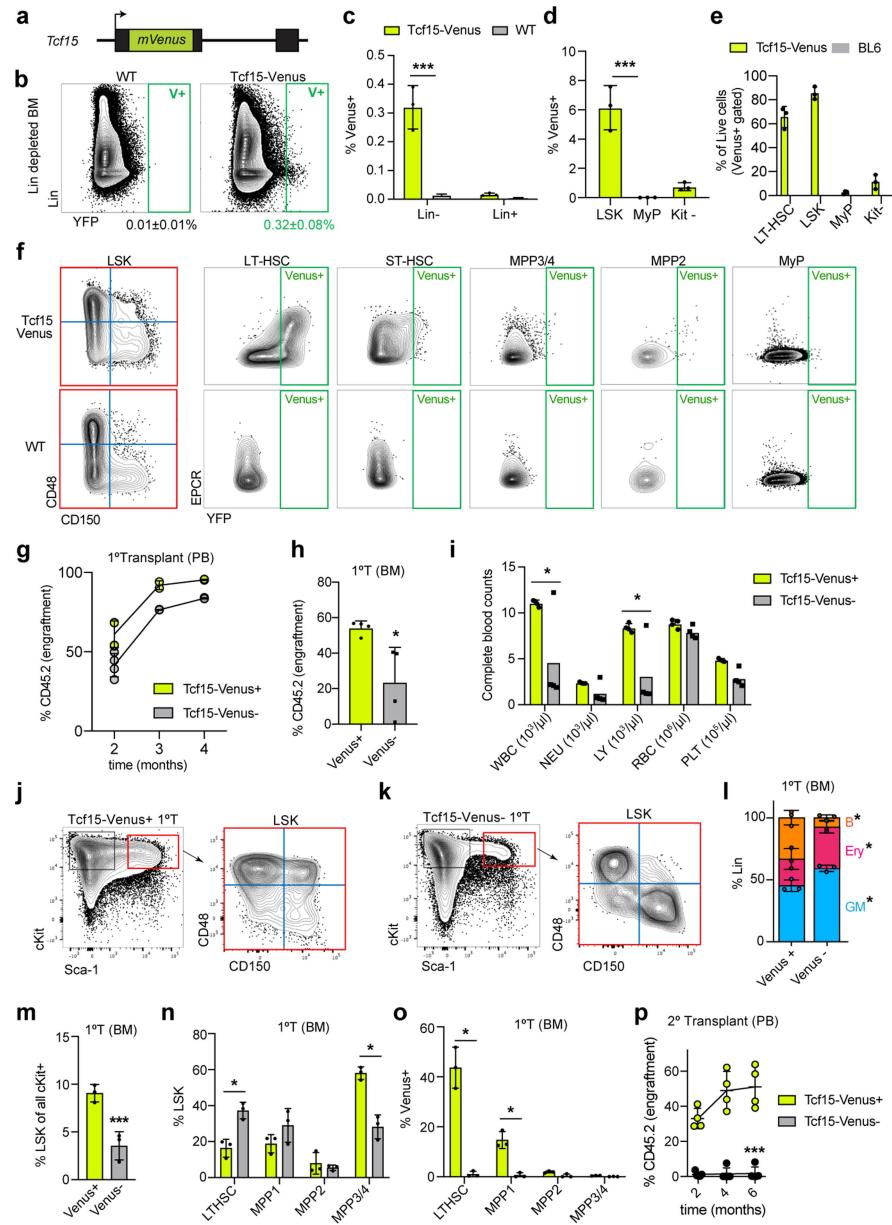
*significant discoveries. $P_{\text{LT-HSC}} < 0.0001$, $P_{\text{MPP1}} = 0.0237$, $P_{\text{MPP2}} = 0.1427$, $P_{\text{MPP3/4}} = 0.5190$, $P_{\text{Myp}} = 0.1206$, $P_{\text{Mkp}} = 0.5190$, $P_{\text{GM}} = 0.0002$, $P_{\text{preB}} < 0.0001$

(two-sided Holm-Sidak multiple-corrected *t*-test, $n = 3$). **f**, Phenotype quantification as Mean \pm s.d. % of donor LSKs in primary recipients corresponding to each SLAM gate (LT-HSC, MPP1, MPP2, MPP3/4). *significant P -value $P_{\text{LT-HSC}} < 0.0001$, $P_{\text{MPP1}} = 0.0001$, $P_{\text{MPP2}} = 0.7152$, $P_{\text{MPP3/4}} = 0.0428$ (two-sided Holm-Sidak multiple-corrected *t*-test, $n = 3$). **g**, FACS scatter plots of sgControl and sgTcf15 EGFP+ LSKs, stained with DAPI and Ki-67 to evaluate cell cycle status. Plots are taken from representative one animal per group taken from 3 independent experiments.



Extended Data Fig. 8 | Additional data on Tcf15 sufficiency for HSC quiescence. **a**, Micrographs of liquid cultures of control TetO-Tcf15 cells. LT-HSCs (1000 cells) from M2rtTA mice were transduced with GFP-carrying lentiviral vectors expressing either a control sgRNA or TetO-Tcf15. Cells were sorted immediately into 1 µg/ml Dox-supplemented STEMspan + SCF/Flt3L/TPO and cultured for 7 days. Images are representative of 5 independent experiments with similar results. **b**, Quantification of liquid culture cellularity by measuring the area of the liquid colonies from 5 independent experiments. Mean ± s.d. is indicated. Control HSC cultures are shown in black, and TetO-Tcf15 HSC cultures are shown in green. * $P < 0.0001$ (unpaired two-sided *t*-test). **c**, Experimental setup to evaluate the effect of Tcf15 overexpression. **d**, Quantification of TetO-Tcf15 EGFP+ cells in peripheral blood. Time-point 0 reflects the lentiviral transduction efficiency evaluated from a remainder of non-transplanted cultured HSCs. Untreated (Dox-) controls ($n=5$) were compared with Dox-treated (Dox+) mice ($n=5$). Line represents mean. Arrow indicates time point of Dox addition in the Dox-treated mice. *** Two-way ANOVA test (genotype x time-factor) $P=0.0127$. **e**, FACS contour plots of Dox-treated TetO-Tcf15 bone marrow cells at 16 wk. Left panels show Lin-EGFP-control cells. Right panels show Lin-EGFP+ TetO-Tcf15 cells. Plots are representative from 3 independent experiments. **f**, Fraction of TetO-Tcf15 EGFP+ cells in different bone marrow populations at 16wk ($n_{\text{Dox}}=5$, $n_{\text{Dox}+}=3$). Mean ± s.d. *two-sided unpaired *t*-test. P-values are $P_{\text{LT-HSC}}=0.0144$,

$P_{\text{MyP}}=0.0010$, $P_{\text{GM}}=0.0091$, $P_{\text{preB}}=0.0032$. **g**, Quantification of % of all Lin-EGFP+ cells that belong to the LT-HSC or MPP1(ST-HSC) fraction ($n_{\text{Dox}}=5$, $n_{\text{Dox}+}=3$). Mean ± s.d. *two-sided unpaired Holm-Sidak-corrected multiple comparisons *t*-test. P-values are $P_{\text{LT-HSC}}=0.0062$, and $P_{\text{MPP1}}=0.0157$. **h**, Quantification of LT-HSC, MPP1, MPP2 and MPP3/4 as % of all donor LSK, comparing EGFP+ (treated and untreated) and EGFP- cells ($n_{\text{Dox}}=5$, $n_{\text{Dox}+}=3$). Mean ± s.d. *two-sided unpaired Holm-Sidak-corrected multiple comparisons *t*-test. P-values are $P_{\text{LT-HSC}}=0.0042$, and $P_{\text{MPP3-4}}=0.0001$. **i**, Quantification of cell cycle phase (G0, G1, G2/M) in LT-HSCs, comparing donor EGFP+ (Dox-treated and untreated) and EGFP- cells ($n_{\text{Dox}}=5$, $n_{\text{Dox}+}=3$). Mean ± s.d. *two-sided unpaired Holm-Sidak-corrected multiple comparisons *t*-test, $P_{\text{G0}}=0.0148$, $P_{\text{G1}}=0.1127$, $P_{\text{G2/S/M}}=0.4815$. **j**, Competitive secondary transplantation of cKit cells derived from Dox-supplemented TetO-Tcf15 mice. EGFP+ cKit+ cells were FACS-purified from Dox-treated primary recipients from experiment in Extended Data Fig. 8c. These cells were transplanted competitively against the same number of cKit cells isolated from a CD45.2+ wild-type donor (same gate), with an additional 250,000 of CD45.1 nucleated whole bone marrow cells (WBM). **k**, Quantification of EGFP+ CD45.2+ secondary engraftment showing higher repopulation from TetO-Tcf15 cKit+ cells (EGFP positive), which outcompete WT cKit+ cells (EGFP negative). Line represents mean ($n=4$ independent experiments). One-way *t*-test (vs. null hypothesis of 50% engraftment) $P=10^{-202}$.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Additional data on the *Tcf15*-Venus knock-in mouse model. **a**, *Tcf15*-Venus knock-in mouse allele. The open-reading frame of monomeric Venus fluorescent protein is knocked-in replacing the start codon in the first exon of the *Tcf15* locus. **b**, FACS plot of *Tcf15*-Venus knock-in mouse reporter bone marrow, stained with Lineage markers. Bone marrow from a wild-type BL/6J mouse is used as a negative control. The YFP channel was used to detect expression of Venus fluorescent protein. Plots are representative of 3 independent experiments with similar results. **c**, Quantification of %Venus+ cells in Lin- vs. Lin+ bone marrow, comparing *Tcf15*-Venus reporter and negative control mice ($n=3$). Mean \pm s.d. ***Holm-Sidak-corrected multiple comparison two-sided t -test $P=0.0243$. **d**, Quantification of %Venus+ cells in Lin-Scal+cKit+ (LSK), Lin-Scal-cKit+ (MyP) and Lin-Scal-cKit- (Kit-). Mean \pm s.d. ***unpaired two-sided t -test, $P=0.0021$ ($n=3$). **e**, Quantification of distribution of Lin- Venus+ cells from *Tcf15*-Venus knock-in reporter bone marrow (measured as % Live Lin-). BL/6J bone marrow cells are shown for comparison, as negative controls. Mean \pm s.d. ($n=3$). **f**, FACS plot of *Tcf15*-Venus knock-in reporter LSK cells, stained for LSK SLAM markers to show YFP (Venus) expression in different SLAM compartments. BL/6J bone marrow LSK cells are used as a negative control. Plots shown are representative of 3 independent experiments with similar results. **g**, Donor engraftment in primary competitive transplantation, measured as % of peripheral blood CD45.2⁺ leukocytes. Bars indicate mean \pm s.d. ($n=4$). **h**, Engraftment in bone marrow, measured as total CD45.2⁺ cells at 3–4 months post transplantation. Mean \pm s.d. ($n=4$). *Holm-Sidak-corrected multiple comparison unpaired two-sided t -test, $P=0.0223$. **i**, Automated peripheral blood counts of mice reconstituted with Venus+ or Venus- HSCs. The scale is shared for all measurements, but the units are

indicated for each population after the labels. *Holm-Sidak-corrected multiple comparison two-sided t -test $P_{WBC}=0.0006$, $P_{LY}=0.0056$. **j**, FACS plots showing bone marrow Lin- analysis of primary recipients transplanted with Venus+ HSCs. Left panels show cKit vs. Scal staining of all cKit⁺ cells. Right panel shows SLAM (CD48, CD150) staining of LSK cells. Plots shown are representative of 3 independent experiments with similar results. **k**, FACS plots showing bone marrow Lin- analysis of primary recipients transplanted with Venus- HSCs. Left panels show cKit vs. Sca1 staining of all cKit⁺ cells. Right panel shows SLAM (CD48, CD150) staining of LSK cells. Plots shown are representative of 3 independent experiments with similar results. **l**, Quantification of % of bone marrow myeloid (GM, Gr-1⁺), lymphoid (B, CD19⁺) and erythroid (Ery, Ter119⁺) cells from Venus+ vs. Venus- primary recipients. Mean \pm s.d. ($n=3$). *Holm-Sidak corrected multiple comparison two-sided t -test. $P_B=0.0002$, $P_{Ery}=0.0166$, $P_{GM}=0.0125$. **m**, Quantification of FACS gate in (j, left panels) showing % of all cKit cells that are LSK. Mean \pm s.d. ($n=3$). ***unpaired two-sided t -test. $P_B=0.0054$. **n**, Quantification of % of donor-derived LSK cells belonging to each SLAM population. Mean \pm s.d. ($n=3$). *Holm-Sidak corrected multiple comparison two-sided t -test. $P_{LT-HSC}=0.0010$, $P_{MPP1}=0.0806$, $P_{MPP2}=0.6026$, $P_{MPP3-4}<0.0001$. **o**, Quantification of % Venus+ cells in each CD45.2⁺ LSK SLAM subpopulation, comparing recipients transplanted with 100 Venus+ vs. Venus- HSCs. Mean \pm s.d. ($n=3$). *Holm-Sidak corrected multiple comparison two-sided t -test. $P_{LT-HSC}<0.0001$, $P_{MPP1}=0.0002$, $P_{MPP2}=0.8157$, $P_{MPP3-4}=0.8820$. **p**, Donor engraftment in secondary competitive transplantation, measured as % of peripheral blood CD45.2⁺ granulocytes. Mean \pm s.d. ($n_{Venus+}=4$, $n_{Venus-}=5$). Line connects the means at each time point. ***paired two-sided t -test. $P<0.0001$.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Single cell mRNASeq and DNASeq libraries were performed with inDrops (v3) as per the guidelines of the Harvard Medical School Single Cell Core. Libraries were sequenced at Harvard Biopolymers facility, using Illumina NextSeq500, and requesting BCL files. BCL files were processed with Illumina Bcl2fastq (v2.20) to generate the R1-R2-R3-R4 files that are required for the inDrops pipeline. FACS data was collected using BD FACS software (FACSDiva v8), and analyzed using FlowJo v10.

Data analysis

Single cell RNASeq analysis used [github.com/indrops \(v0.3\)](https://github.com/indrops/v0.3) and [github.com/AllonKleinLab/LARRY \(v0.1\)](https://github.com/AllonKleinLab/LARRY (v0.1)). For analysis and visualization, we used SPRING: [github.com/AllonKleinLab/SPRING_dev \(v1.6\)](https://github.com/AllonKleinLab/SPRING_dev (v1.6)) and Scanpy: [github.com/theislab/scanpy \(v1.4.6\)](https://github.com/theislab/scanpy (v1.4.6)). For inspecting cell doublets, we used SCRUBLET: [github.com/AllonKleinLab/Scrublet \(v0.1\)](https://github.com/AllonKleinLab/Scrublet (v0.1)). For single cell CROP-seq experiments, the LARRY script was minimally modified to search for CROPseq sgRNA scaffold sequences, instead of lentiviral LARRY barcodes. For analysis of DNASeq of CROP-seq experiments, reads were mapped using bowtie (v1.2) to a simple bowtie index built from the sgRNA sequences of the CROP-seq library.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data and analyses are available at the following links: http://github.com/rodriguez-fraticelli/Tcf15_HSCs and <https://github.com/AllonKleinLab/StemCellTransplantationModel>. Raw data and counts matrices are available at GEO (GSE134242). Source Data behind Figures 1-4 and Extended Data Figures 1-9 are available within the manuscript files.

The LARRY barcoding tool is available at Addgene (#140024).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The number of samples collected was estimated by prior studies using lentiviral barcoding (see references), but this number was not statistically predefined or optimized. We sought to collect enough samples (clones), in order to have enough cells of different categories to allow further analysis of differential gene expression.
Data exclusions	No data was excluded.
Replication	The number of replicates is indicated in each experiment. Experiments were successfully and independently replicated at least 3 times, and usually more than 4 times, with similar results, as indicated in each figure. Single cell mRNASeq experiments (with the exception of TetO-Tcf15) were replicated at least 2 times, with similar results. The most important conclusions of single cell experiments were all validated by flow-cytometry.
Randomization	In all cases, donors and recipients of the transplants were randomly selected from pools of untagged littermates.
Blinding	Blinding was not required in most experiments. Computational analysis required proper identification of clone sources, and this was not suitable for blinding. For animal analyses, samples were required to be collected, labeled and analyzed by the same authorized researchers, complicating blinding study design without bias. Blinding was used only in primary and secondary transplantation of Tcf15-sgRNA and Tcf15-Venus+/- HSCs, to avoid biases during transplantation procedures. This was performed by labeling each sample with a number after cell sorting, covering the number with tape, and then picking tubes randomly before injection. After injection, animal cages were then labeled with the correct genotype.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

- | | |
|-------------------------------------|-----------------------------|
| n/a | Involved in the study |
| <input type="checkbox"/> | Antibodies |
| <input type="checkbox"/> | Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | Palaeontology |
| <input type="checkbox"/> | Animals and other organisms |
| <input checked="" type="checkbox"/> | Human research participants |
| <input checked="" type="checkbox"/> | Clinical data |

Methods

- | | |
|-------------------------------------|------------------------|
| n/a | Involved in the study |
| <input checked="" type="checkbox"/> | ChIP-seq |
| <input type="checkbox"/> | Flow cytometry |
| <input checked="" type="checkbox"/> | MRI-based neuroimaging |

Antibodies

Antibodies used

See Supplementary Table 13 - antibodies.

Antibody / Clone / Manufacturer / Concentration / Catalog / Validation
 Ly6G APC-eFluor780 1A8 eBioscience 1:50 ThermoFisher Scientific #47-9668-82 Y
 Ly6G Alexa Fluor 700 1A8 eBioscience 1:50 ThermoFisher Scientific #56-9668-82 Y
 CD117 (cKit) FITC 2B8 eBioscience 1:100 ThermoFisher Scientific #11-1171-82 Y
 CD117 (cKit) APC 2B8 eBioscience 1:100 ThermoFisher Scientific #17-1171-83 Y
 CD117 (cKit) APC ACK2 Biolegend 1:100 Biolegend #135108 Y
 CD117 (cKit) APC ACK2 eBioscience 1:100 ThermoFisher Scientific # 17-1172-82 Y
 Ly6a/e (Sca-1) PE D7 eBioscience 1:100 ThermoFisher Scientific # 12-5981-82 Y
 Ly6a/e (Sca-1) PE/Cy7 D7 eBioscience 1:100 ThermoFisher Scientific # 25-5981-82 Y
 CD150 PE/Cy5 TC15-12F12.2 Biolegend 1:100 Biolegend #115912 Y
 CD48 APC/Cy7 HM48-1 BD Biosciences 1:100 BD Biosciences #561242 Y
 CD41 BV605 MWReg30 Biolegend 1:100 Biolegend #133921 Y

CD3e biotin 145-2C11 BD Biosciences 1:100 BD Biosciences #553060 Y
 CD3 eFluor450 145-2C11 eBioscience 1:100 ThermoFisher Scientific #48-0031-82 Y
 CD3e PE 145-2C11 eBioscience 1:100 ThermoFisher Scientific #12-0031-81 Y
 CD19 biotin MB19-1 eBioscience 1:100 ThermoFisher Scientific #13-0191-82 Y
 CD19 eFluor450 1D3 eBioscience 1:100 ThermoFisher Scientific #48-0193-80 Y
 CD19 APC/Cy7 1D3 eBioscience 1:100 ThermoFisher Scientific #47-0193-82 Y
 Gr1 biotin RB6-685 eBioscience 1:100 ThermoFisher Scientific #13-5931-82 Y
 Gr1 eFluor450 RB6-685 eBioscience 1:100 ThermoFisher Scientific #48-5931-82 Y
 Gr1 PE/Cy7 RB6-685 eBioscience 1:100 ThermoFisher Scientific #25-5931-82 Y
 Gr1 PE/Cy5 RB6-685 eBioscience 1:100 ThermoFisher Scientific #15-5931-82 Y
 CD11b (Mac1) biotin M1/70 eBioscience 1:100 ThermoFisher Scientific #13-0112-85 Y
 CD11b (Mac1) eFluor450 M1/70 eBioscience 1:100 ThermoFisher Scientific #48-0112-82 Y
 Ter119 biotin TER119 eBioscience 1:100 ThermoFisher Scientific #13-5921-82 Y
 Ter119 eFluor450 TER119 eBioscience 1:100 ThermoFisher Scientific #48-5921-82 Y
 Ter119 PE/Cy5 TER119 eBioscience 1:100 ThermoFisher Scientific #15-5921-82 Y
 Streptavidin eFluor 450 - eBioscience 1:200 ThermoFisher Scientific #48-4317-82 Y
 EPCR PE (Procr) RCR-16 Biolegend 1:50 Biolegend #141504 Y
 Ki67 PE/Cy7 16A8 Biolegend 1:100 Biolegend #652426 Y
 CD45.1 APC A20 Biolegend 1:100 Biolegend #110714 Y
 CD45.1 PE A20 Biolegend 1:100 Biolegend #110708 Y
 CD45.1 BV711 A20 Biolegend 1:100 Biolegend #110739 Y
 CD45.2 APC 104 Biolegend 1:100 Biolegend #109814 Y
 CD45.2 PE 104 Biolegend 1:100 Biolegend #109808 Y
 CD45.2 BV605 104 Biolegend 1:100 Biolegend #109841 Y

Validation

Antibodies are all commercial monoclonal antibodies of well known clones, and specifically tested in flow-cytometry and cell sorting. These commercial antibodies are widely described and reliably tested in the literature. Each antibody has also been validated by the manufacturers for flow-cytometry, including antigen specificity.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	293T (ATCC CRL-3216) cells were used for lentiviral preparation
Authentication	No authentication was performed on this line.
Mycoplasma contamination	Cell lines were regularly (at least 1 every 6 months) tested for mycoplasma contamination by supernatant PCR. In all cases, PCRs were negative.
Commonly misidentified lines (See ICLAC register)	293T. These cell lines are only used for lentiviral packaging. No biological conclusions are derived from their observation in this study.

Animals and other organisms

Policy information about [studies involving animals; ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	The TetO-Cas9/M2rtTA mice (c57BL/6 background) were a kind gift from Stuart Orkin (and are available from The Jackson Laboratory strain #029476). The Tcf15-Venus (c57BL/6 background) mice were generated at the University of Edinburgh. All other mice were BL/6J strain (Ly5.2 and Ly5.1) and obtained from The Jackson Laboratory. 8-10wk old females were used as recipients for all experiments. 8-10wk old males and females were used indistinctly for all other experiments. All animal procedures were approved by the Boston Children's Hospital Institutional Animal Care and Use Committee.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve field-collected samples.
Ethics oversight	The protocol is approved by Boston Children's Hospital Institutional Animal Care and Use Committee, accredited by the AAALAC.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- 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).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Whole bone marrow (WBM) was prepared by crushing, filtered using a 70 um mesh and centrifuged at 750g for 8 minutes. WBM isolates were depleted of red blood cells using RBC-lysis buffer, and then samples were washed, filtered through a 40 um mesh and centrifuged again at 750g for 7 minutes. Then, cells were either lineage-depleted or KIT-enriched using magnetic separation columns (Miltenyi), and then stained with the indicated antibodies as per the instructions of the manufacturers. 1% Pen-Strep, 2% Fetal Bovine Serum-supplemented cold PBS was used for all isolation steps and incubations. Cells were resuspended in 1% PS-2%FBS PBS with 1 ug/ml DAPI (dead-cell stain) before flow cytometry. For sorting, cells were filtered again using a 40um-filter-cap 5 ml tube before sorting, and then sorted into 1ml of 2%FBS-supplemented PBS in 1.5 ml centrifuge tubes.

Instrument

BD Aria IIu and BD Aria Fusion were used for fluorescence-activated cell sorting. BD LSRII and Fortessa were used for flow cytometry.

Software

FACS data was collected using BD FACS software (FACSDiva), and analyzed using FlowJo v10.

Cell population abundance

Samples were sorted at 1500 events/sec with a 4-way purity mode, and using the 85um nozzle, achieving ~95% purity by FACS analysis. Efficiency was typically 80-90%.

Gating strategy

The following combinations of cell surface markers were used to define these cell populations: Erythroblasts: Ly6G- CD19- Ter119+ FSChi, Granulocytes: Ly6G+ CD19- Ter119-, Monocytes: Ly6C+ Ly6G- CD19- Ter119-, pro/pre-B cells: Ly6G- CD19+, Megakaryocyte progenitors: Lin- cKit+ Sca1- CD150+ CD41+, LT-HSC: Lin- cKit+ Sca1+ CD150+ CD48- MPP1/ST-HSC: Lin- cKit+ Sca1+ CD150- CD48-, MPP2: Lin- cKit+ Sca1+ CD150+ CD48+, MPP3/4: Lin- cKit+ Sca1+ CD150- CD48+.

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