Population snapshots predict early haematopoietic and erythroid hierarchies

Betsabeh Khoramian Tusi1*, Samuel L. Wolock2*, Caleb Weinreb2*, Yung Hwang1, Daniel Hidalgo1, Rapolas Zilionis2, Ari Waisman3, Jun R. Huh4, Alon M. Klein2 & Merav Socolovsky1,5

The formation of red blood cells begins with the differentiation of multipotent haematopoietic progenitors. Reconstructing the steps of this differentiation represents a general challenge in stem-cell biology. Here we used single-cell transcriptomics, fate assays and a theory that allows the prediction of cell fates from population snapshots to demonstrate that mouse haematopoietic progenitors differentiate through a continuous, hierarchical structure into seven blood lineages. We uncovered coupling between the erythroid and the basophil or mast cell fates, a global haematopoietic response to erythroid stress and novel growth factor receptors that regulate erythropoiesis. We defined a flow cytometry sorting strategy to purify early stages of erythroid differentiation, completely isolating classically defined burst-forming and colony-forming progenitors. We also found that the cell cycle is progressively remodelled during erythroid development and during a sharp transcriptional switch that ends the colony-forming progenitor stage and activates terminal differentiation. Our work showcases the utility of linking transcriptomic data to predictive fate models, and provides insights into lineage development in vivo.
distribution of cell states. PBA approximates the dynamics of cells through ongoing cell turnover, explains the observed steady-state reconstruction of the memoryless stochastic dynamics of cells, which, been detailed elsewhere33. The core of PBA can be understood as the cell fate probabilities. The derivation and limitations of PBA have space that encodes the cell-graph topology in the form of predicted (Extended Data Fig. 1c, d). PBA maps each cell to a low-dimensional space that encodes the putative cell fate probabilities. The probabilities obtained by PBA represent formal biophysical predictions for the fate of cells under simplified assumptions, but they can also be treated heuristically as encoding graph distances. Applied to our data, PBA defines seven putative commitment probabilities for each haematopoietic progenitor (Fig. 1c, Extended Data Fig. 1e), as well as the distance from the undifferentiated CD34highSca1high MPPs (Fig. 1d).

The transcriptional continuum of HPCs is hierarchical

We used PBA-predicted commitment probabilities to compute a coupling score that reflects whether any two fate potentials occur concurrently in single progenitors at higher rates than would be expected by chance (Fig. 1e). A transcriptional-state hierarchy was formalized by identifying correlated pairs of terminal fates and joining them iteratively until a multipotent state was reached (Fig. 1e, f). The resulting topology firmly supports the hierarchical view of haematopoiesis, with MPPs diverging into progenitors with either correlated erythroid, Ba/Mast and megakaryocytic fates, or with correlated lymphoid and myeloid fates (Fig. 1f). However, the transcriptional-state hierarchy emerges from correlations on a continuum, rather than from discrete populations. Additionally, it predicts two refinements over current models. First, the erythroid fate is correlated with the Ba/Mast fates. Second, among myeloid progenitors we identified dendritic–monocyte and granulocytic–monocyte coupling, but no dendritic–granulocytic coupling, suggesting that monocyte differentiation may occur through two distinct trajectories, a prediction that was very recently independently confirmed34. The PBA-formalized HPC hierarchy also allowed us to identify genes for which expression closely correlates with each cell fate choice (Extended Data Fig. 2, Supplementary Table 2).

Isolation of putative erythroid progenitors

To test PBA predictions (Fig. 1e–f), we developed a FACS strategy that isolates haematopoietic subpopulations defined by scRNA-seq. Guided by the single-cell expression patterns, we combined Kit expression with CD55, a marker of megakaryocytic and erythroid bias10; Kit+ CD55+ cells were divided into subpopulations (P1–P5), using CD49f (encoded by *Itga6*) and megakaryocytic and erythroid markers6,9,10 (Fig. 2a). Using reverse transcription with quantitative PCR (qRT–PCR) (Extended Data Fig. 3a), and scRNA-seq (11,241 cells post-filter) (Fig. 2b, Extended Data Fig. 3b–e), we mapped cells from each of the sorted subpopulations back to regions of the SPRING graph. We found that P1 and P2 represent high-purity subpopulations on the putative erythroid branch, with P1 predicted to be committed, and P2 mostly committed, to the erythroid fate (Fig. 1c). P3 and P4 are enriched for the Ba/Mast and megakaryocytic branches, respectively; P3 bifurcates into separate basophil and mast cell branches. P5 contains erythroid-biased oligopotent and multipotent cells. Myeloid, lymphoid and some MPPs are within the CD55− region of the plot.

Functional identification of correlated cell fates

We next examined the differentiation potential of the sorted P1–P5 populations, and by extension, the predicted fate probabilities for their corresponding transcriptional states. Colony-forming assays showed that P1 and P2 contain all of the unipotential erythroid progenitors and no other progenitors (Fig. 2c–e). P1 colonies were small and unifocal, maturing on day 3 or later (CFU-e) (Fig. 2c), whereas P2 colonies were largely multifocal, maturing on day 4 or later (BFU-e) (Fig. 2d). Thus, P1 is closer to erythroid maturation than P2, consistent with PBA predictions (Figs 1c, 2b). Furthermore, the transcriptional state of progenitors as defined by the SPRING map determines their ability to form either multifocal or unifocal colonies. Consistent with the SPRING plot, the less-differentiated P5 population gave rise to mixed myeloid colonies, and P4 was enriched for megakaryocytic progenitors (Fig. 2e, Extended Data Fig. 4a).
To test HPC fate potential further, we sorted single Kit⁺ cells into liquid culture wells in the presence of cytokines that support myeloid and erythroid differentiation (Fig. 3a). We assayed the clonal output of 1,158 single cells by FACS (Fig. 3b, Extended Data Fig. 4b). Unipotential clones for the erythroid, Ba/Mast, megakaryocytic, and granulocytic/monocytic (GN/M) lineages largely originated in the P1/P2, P3, P4 and CD55⁺ subpopulations, respectively, consistent with predictions (Figs 1c, 2b, 3b). Many clones contained multiple lineages, with strong, statistically significant couplings between the erythroid, Ba/Mast and megakaryocytic cell fates on the one hand, and the GN/M fates on the other (Fig. 3c; absolute z-score value of more than 10 when compared to randomized data), consistent with both known (erythroid and megakaryocytic, GN/M) and novel (erythroid and Ba/Mast) PBA predictions (Fig. 1c, e, f). Progenitors with erythroid and Ba/Mast output were enriched in the P2 and P5 subpopulations (Fig. 3d), which map close to the erythroid and Ba/Mast branch point in the scRNA-seq data (Figs 1c, 2b), and were depleted in the CD55⁻ population, as predicted. We found similar results in bulk liquid cultures (Extended Data Fig. 4c). Notably, the new Ba/Mast differentiation pathway suggested by our data does not rule out Ba/Mast formation by the traditional route, akin to published pseudotemporal-ordering algorithms35–37 (Fig. 4b). Notably, the new Ba/Mast differentiation pathway suggested by our data does not rule out Ba/Mast formation by the traditional route, akin to published pseudotemporal-ordering algorithms35–37 (Fig. 4b).}

To test HPC fate potential further, we sorted single Kit⁺ cells into liquid culture wells in the presence of cytokines that support myeloid and erythroid differentiation (Fig. 3a). We assayed the clonal output of 1,158 single cells by FACS (Fig. 3b, Extended Data Fig. 4b). Unipotential clones for the erythroid, Ba/Mast, megakaryocytic, and granulocytic/monocytic (GN/M) lineages largely originated in the P1/P2, P3, P4 and CD55⁺ subpopulations, respectively, consistent with predictions (Figs 1c, 2b, 3b). Many clones contained multiple lineages, with strong, statistically significant couplings between the erythroid, Ba/Mast and megakaryocytic cell fates on the one hand, and the GN/M fates on the other (Fig. 3c; absolute z-score value of more than 10 when compared to randomized data), consistent with both known (erythroid and megakaryocytic, GN/M) and novel (erythroid and Ba/Mast) PBA predictions (Fig. 1c, e, f). Progenitors with erythroid and Ba/Mast output were enriched in the P2 and P5 subpopulations (Fig. 3d), which map close to the erythroid and Ba/Mast branch point in the scRNA-seq data (Figs 1c, 2b), and were depleted in the CD55⁻ population, as predicted. We found similar results in bulk liquid cultures (Extended Data Fig. 4c). Notably, the new Ba/Mast differentiation pathway suggested by our data does not rule out Ba/Mast formation by the traditional route, akin to published pseudotemporal-ordering algorithms35–37 (Fig. 4b). Notably, the new Ba/Mast differentiation pathway suggested by our data does not rule out Ba/Mast formation by the traditional route, akin to published pseudotemporal-ordering algorithms35–37 (Fig. 4b).
A further, sharp induction of Tfrc takes place at the transition to ETD (Fig. 4c).

Of the approximately 4,500 genes that varied significantly along the erythroid trajectory (Supplementary Table 3), a large group was induced at the onset of the CEP stage, and sharply suppressed at the CEP to ETD transition (Fig. 4b). It contained the most dominant module. Our analysis predicts new epigenetic and transcriptional pathways that are conserved between species. These pathways suggest that CEPs, which are least (0%) to the most (100%) differentiated.

**Erythroid stress generates a broad response**

Using scRNA-seq, we examined two model systems of accelerated, or stress, erythropoiesis: the mid-gestation fetal liver (n = 7,182 cells post-filter), in which erythropoiesis is a rate-limiting factor of fetal growth; and bone marrow from mice treated with EPO for 48 h, stimulating red blood cell production (n = 2,611 cells post-filter). SPRING graphs revealed a remarkable conservation of the key features of the haematopoietic hierarchy and erythroid differentiation during stress (Fig. 4d). The proportion of erythroid-trajectory cells increased with stress (Fig. 4e, f). In the fetal liver, the increase was predominantly in CEPs, whereas in EPO-stimulated bone marrow, all erythroid-trajectory cells increased in abundance, including uncommitted MPPs and EBMPs. We found that the absolute number of Kit+ cells in EPO-stimulated bone marrow did not change, indicating that the increase in erythroid-trajectory cells came at the expense of other cell lineages (Fig. 4e, Extended Data Fig. 7). A number of mechanisms could account for this, including altered intrinsic fate bias of MPPs.39,40

EPO addition altered gene expression principally in EEPs and CEPs, but also in EBMPs and MPPs (Fig. 4f), in which targets of realised...
Figure 5 | Novel growth factor regulators of early erythropoiesis.

a, Expression patterns for Mst1r, Ryk and Il17ra (see Extended Data Fig. 9a, b). TPM, transcripts per million. b, Effect of MSP, WNT5A or IL-17A on EPO-dependent CFU-e colony formation. Bars represent the mean of two or three independent experiments (individual data points), each performed in quadruplicate (for full analysis see Extended Data Fig. 9c). c, The IL-17A response is lost in Il17ra−/− bone marrow. Data are mean ± s.d. per 500,000 bone marrow cells plated in triplicate in the presence of EPO (0.05 U ml−1) and are representative of two independent experiments. d, IL-17A stimulates CFU-e formation in freshly isolated human bone marrow mononuclear cells. Data are mean ± s.d. per 85,000 cells plated in triplicate. e, IL-17A-mediated phosphorylation of STAT3 and STAT5 (pSTAT3 and pSTAT5). Fresh bone marrow cells were starved of cytokines for 3 h, and then stimulated with EPO, IL-17A or both; FACs profiles are for baseline (starved, shown in grey), and 60 min after stimulation (in colour). Profiles are representative of two independent experiments, each performed in duplicate. FU, fluorescence units. f, Summary of growth factor effects on erythroid output.

CCAAT-enhancer-binding protein β (C/EBPβ), a transcription factor that biases differentiation away from erythroid and megakaryocytic fates, were downregulated. We identified both known and new stress-responsive genes, together with their precise localization within the erythroid trajectory (Extended Data Fig. 8, Supplementary Table 5).

Taken together, we found that the cell-state branching structure is maintained during accelerated erythropoiesis. In MPPs and throughout the ensuing erythroid progression, we identified changes in gene expression and in cell abundance in response to EPO well beyond the currently known mechanism of EPO-driven erythropoietic expansion.

Growth factor receptors of early erythropoiesis

We screened EEPs and CEps for gene expression of cell-surface receptors with known ligands using qRT–PCR, identifying three such receptors encoded by Ryk, Mst1r and Il17ra (Fig. 5a, Extended Data Fig. 9a, b). Ryk and Mst1r have previously been reported in CFU-e, but their function remains unknown. However, the expression of an IL-17A receptor by EEPs has not, to our knowledge, been documented. We stimulated Ryk, Mst1r and IL-17A with their respective ligands, WNT5A, macrophage-stimulating protein (MSP) and IL-17A, using erythroid colony formation as a readout (Fig. 5b, Extended Data Fig. 9c). In the fetal liver sample, in the presence of low levels of EPO (50 mU ml−1), MSP doubled the number of CFU-e colonies, equivalent to a tenfold increase in EPO concentration. MSP was inhibitory in other contexts, and WNT5A was a potent inhibitor of erythroid colony formation in both the fetal liver and bone marrow samples. By contrast, IL-17A mediated a marked potentiation of adult bone marrow CFU-e colony formation, quadrupling colonies at low levels of EPO (50 mU ml−1), and increasing them by approximately 50% at high levels of EPO (500 mU ml−1).

The stimulatory effect of IL-17A required endogenous IL-17RA (Fig. 5c) and was also evident in human bone marrow (Fig. 5d). Furthermore, IL-17A stimulation was saturable, with a low half-maximum effective concentration (EC50) (60 pM), consistent with high-affinity binding of IL-17A to IL-17RA. IL-17A induced rapid phosphorylation of the intracellular signalling mediators STAT3 and STAT5 in CEps and EEPs (Fig. 5e), and western blotting of freshly sorted CEp P1 and EEP P2 cells revealed expression of IL-17RA (Extended Data Fig. 9d). Taken together, our findings suggest previously unknown regulation of CEps and CEps through the expression of a number of growth factor receptors.

Cell cycle remodelling during erythroid development

In a final analysis, we asked what governs progression through the CEp stage and its termination during ETD. We previously reported that the onset of ETD in the fetal liver occurs within a single S phase, and is dependent on S-phase progression; furthermore, this unique S phase is shorter and faster than the S phase in pre-ETD cells. These conclusions, which are based on the analysis of large fetal liver subpopulations, predict that CEP exit should show an S phase signature. In our scRNA-seq data, we found that the expression levels of genes that mark the G1/S, S, G2 and G2/M cell cycle phases form a sequence of close, sharp peaks during CEP exit, probably representing a single cell cycle (Fig. 6a, b). This and the following results hold even when cell cycle genes are omitted for ordering the erythroid trajectory (Extended Data Fig. 10a–c). Notably, by reversibly inhibiting DNA replication, we found that the CEP-to-ETD transition in adult bone marrow not only synchronized with, but also depended on, S phase progression (Extended Data Fig. 10d–f).

The scRNA-seq data revealed that changes to cell cycle machinery occur throughout the CEp stage, perhaps in preparation for the switch to ETD. Genes with expression levels that most closely correlate with CEP progression (Supplementary Table 6) are enriched for Gene Ontology terms associated with DNA replication. Notably, regulators of S phase and the G1/S phase transition increased steadily through the CEP stage, including cyclin E1 (encoded by Ccne1), cyclin A2 (Ccna2) and MCM helicase subunits (Mcm2–Mcm7). Conversely, regulators of the G1 phase such as cyclin D2 (Ccn2) and cyclin-dependent kinase 6 (Cdk6) decrease steadily (Fig. 6c).

To investigate these findings, we labelled S-phase cells in vivo with the nucleotide analogue BrdU, and analysed the cell cycle distribution of cells as they progressed through the EEP and CEp stages (Fig. 6d–f). We found a graded but notable increase in the fraction of cells in S phase, whereas the number of G1 cells correspondingly decreased. Results were similar in both EPO-stimulated bone marrow and fetal liver samples (Extended Data Fig. 10g). There was no significant change in the length or speed of S phases, as evidenced by stable intra-S-phase levels of BrdU (Fig. 6f), suggesting that cells spend more time in S phase as a result of G1 shortening.

Western blotting of sorted P1 and P2 fractions confirmed that the expression of key S-phase regulators increased with development in EEPs and CEps (Extended Data Fig. 10a–c). Notably, by reversibly inhibiting DNA replication, we found that the CEP-to-ETD transition in adult bone marrow not only synchronized with, but also depended on, S phase progression (Extended Data Fig. 10d–f).

In a final analysis, we asked what governs progression through the CEp stage and its termination during ETD. We previously reported that the onset of ETD in the fetal liver occurs within a single S phase, and is dependent on S-phase progression; furthermore, this unique S phase is shorter and faster than the S phase in pre-ETD cells. These conclusions, which are based on the analysis of large fetal liver subpopulations, predict that CEP exit should show an S phase signature. In our scRNA-seq data, we found that the expression levels of genes that mark the G1/S, S, G2 and G2/M cell cycle phases form a sequence of close, sharp peaks during CEP exit, probably representing a single cell cycle (Fig. 6a, b). This and the following results hold even when cell cycle genes are omitted for ordering the erythroid trajectory (Extended Data Fig. 10a–c). Notably, by reversibly inhibiting DNA replication, we found that the CEP-to-ETD transition in adult bone marrow not only synchronized with, but also depended on, S phase progression (Extended Data Fig. 10d–f).

The scRNA-seq data revealed that changes to cell cycle machinery occur throughout the CEp stage, perhaps in preparation for the switch to ETD. Genes with expression levels that most closely correlate with CEP progression (Supplementary Table 6) are enriched for Gene Ontology terms associated with DNA replication. Notably, regulators of S phase and the G1/S phase transition increased steadily through the CEP stage, including cyclin E1 (encoded by Ccne1), cyclin A2 (Ccna2) and MCM helicase subunits (Mcm2–Mcm7). Conversely, regulators of the G1 phase such as cyclin D2 (Ccn2) and cyclin-dependent kinase 6 (Cdk6) decrease steadily (Fig. 6c).

To investigate these findings, we labelled S-phase cells in vivo with the nucleotide analogue BrdU, and analysed the cell cycle distribution of cells as they progressed through the EEP and CEp stages (Fig. 6d–f). We found a graded but notable increase in the fraction of cells in S phase, whereas the number of G1 cells correspondingly decreased. Results were similar in both EPO-stimulated bone marrow and fetal liver samples (Extended Data Fig. 10g). There was no significant change in the length or speed of S phases, as evidenced by stable intra-S-phase levels of BrdU (Fig. 6f), suggesting that cells spend more time in S phase as a result of G1 shortening. Western blotting of sorted P1 and P2 fractions confirmed that the expression of key S-phase regulators increased with development in EEPs and CEps (Extended Data Fig. 10a–c). Notably, by reversibly inhibiting DNA replication, we found that the CEP-to-ETD transition in adult bone marrow not only synchronized with, but also depended on, S phase progression (Extended Data Fig. 10d–f).

Discussion

Our scRNA-seq analysis reveals that HPCs occupy a continuum of transcriptional cell states, branching towards seven fates. Certain cell fate potentials are correlated, supporting a hierarchical view of haematopoiesis, with MPPs diverging either towards myeloid and
We delineated the continuous differentiation trajectory of the erythroid lineage, from its origins in MPPs, through EBMPs, to unipotential EEPs and CEPs, which we show correspond to the unipotential BFU-E and CFU-E, respectively. The dominant CEP stage expresses a distinct transcriptional program and is probably a regulator of erythroid output, as evidenced by both its expansion under stress, and by novel growth factor receptors that regulate CFU-E numbers. In particular, our finding of strong stimulation by the pro-inflammatory IL-17RA contributes to the growing evidence of a complex interplay between erythropoiesis and inflammation.\(^{50,51}\) We further identified the cell cycle as a key process in both the progression and termination of the CEP stage. Developing CEPs spend an increasing fraction of their time in S phase, as a result of G1 shortening; their transition to ETD in an abrupt transcriptional switch is dependent on a single, short S phase. We speculate that the cell cycle may set the context for activation of transcription factors that are induced earlier in the erythroid trajectory. Taken together, our single-cell approach allowed us to make detailed predictions that we validated to reveal novel fundamentals of early hematopoietic differentiation, as well as practical methods for further isolation and study of these cells.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 October 2016; accepted 11 January 2018.

Published online 21 February 2018.

Figure 6 | Extensive remodelling of the cell cycle during erythroid development. a, Cell cycle phase-specific genes\(^{25}\), ordered by peak expression, reveal cell cycle synchronization with the CEP to ETD transition (indicated by an asterisk). b, Mean expression of all genes specific to each cell cycle phase (as in a), traced along the erythroid trajectory. The transition to ETD is marked by a sharp induction of Tfrc. c, Representative cell cycle genes that are correlated or anti-correlated with progression along the erythroid trajectory. d, Schematic for cell cycle analysis of erythroid progenitors in vivo. Bone marrow was collected and fixed 30 min after BrdU injection; P1 and P2 cells were analysed for BrdU incorporation and DNA content. e, BrdU-labelled S-phase cells, as in d. f, CD71 expression (top), cell cycle phase distribution (middle), and intra-S-phase DNA synthesis rate (bottom), for all gates in e. Insets show representative FACS plots of cell cycle distribution. Data are representative of three independent experiments. For similar analyses of EPO-stimulated bone marrow and fetal laval samples, see Extended Data Fig. 10g. g, Summary of cell cycle remodelling during early erythropoiesis and the S-phase-dependent switch to ETD.

lymphoid fates, or towards the erythroid, megakaryocyte and Ba/Mast cell fates. Yet unlike the classical models of haematopoiesis, HPCs do not separate into discrete and homogenous stages. The coupling of specific cell fates, which we validated with single-cell fate assays, is a critical feature by which our model differs from recent models of haematopoiesis, in which unilineage progenitors arise directly from MPPs. Our model also explains historical hierarchical interpretations of haematopoiesis, which were based on fate assays of FACS-gated populations, averaging the fate couplings of their constituent progenitors. Of note, the continuous nature of the scRNA-seq data does not rule out the existence of discrete epigenetic or signalling states among HPCs, if their lifetime in single cells is comparable to, or shorter than, the lifetime of mRNA molecules (in the range of hours to approximately 1 day).

\(^{1}\) Fujisawa, Y., Browne, C. P., Cunniff, K., Goff, S. C. & Orkin, S. H. Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. Proc. Natl Acad. Sci. USA 93, 12355–12358 (1996).

\(^{2}\) Liu, Y. et al. Suppression of Fas–FasL coexpression by erythropoietin mediates erythroblast expansion during the erythropoietic stress response in vivo. Blood 108, 123–133 (2006).

\(^{3}\) Chan, K. et al. Resolving the distinct stages in erythroid differentiation based on dynamic changes in membrane protein expression during erythropoiesis. Proc. Natl Acad. Sci. USA 106, 17413–17418 (2009).

\(^{4}\) Hara, H. & Ogawa, M. Erythropoietin precursors in mice under erythropoietic stimulation and suppression (e.g. Hematol. 5, 141–148 (1977).

\(^{5}\) Gregory, C. J., McCulloch, E. A. & Till, J. E. The cellular basis for the defect in haemopoiesis in flexed-tailed mice. III. Restriction of the defect to erythropoietic progenitors capable of transient colony formation in vivo. Br. J. Haematol. 30, 401–410 (1975).

\(^{6}\) Prong, C. J. et al. Elucidation of the phenotypic, functional, and molecular topography of a myeloperoxygenator progenitor cell hierarchy. Cell Stem Cell 1, 428–442 (2007).

\(^{7}\) Pajegare, J., Rayon Estrada, V., Shin, C., Gupta, S. & Lodish, H. F. HIF1α synergizes with glucocorticoids to promote BFU-E progenitor self-renewal. Blood 117, 3435–3444 (2011).

\(^{8}\) Li, J. et al. Isolation and transcriptome analyses of human erythroid progenitors: BFU-E and CFU-E. Blood 124, 3636–3645 (2014).

\(^{9}\) Mori, Y., Chen, J. Y., Puivinage, J. V., Seita, J. & Weissman, I. L. Prospective isolation of human erythroid lineage-committed progenitors. Proc. Natl Acad. Sci. USA 112, 9638–9643 (2015).

\(^{10}\) Guo, G. et al. Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. Cell Stem Cell 13, 492–505 (2013).

\(^{11}\) Sun, J. et al. Clonal dynamics of native haematopoiesis. Nature 514, 322–327 (2014).

\(^{12}\) Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. Cell 163, 1663–1677 (2015).

\(^{13}\) Busch, K. et al. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. Nature 518, 542–546 (2015).

\(^{14}\) Notta, F. et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. Science 351, 2216 (2016).

\(^{15}\) Nestorowa, S. et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood 128, e20–e31 (2016).

\(^{16}\) Vilella, L. et al. Human haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. 19, 271–281 (2017).

\(^{17}\) Mercier, F. E. & Scadden, D. T. Not all created equal: lineage hard-wiring in the hierarchy across ontogeny. Science 351, 2216 (2016).

\(^{18}\) Kondo, M., Weissman, I. L. & Akashi, K. Identification of clonogenic common lymphoid progenitors in the murine bone marrow. Cell 91, 661–672 (1997).

\(^{19}\) Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. Nature 404, 193–197 (2000).

\(^{20}\) Adolfsson, J. et al. Identification of FI3I–lympho-mycytopoietic stem cells lacking erythroid-megakaryocytic potential a revised road map for adult blood lineage commitment. Cell 121, 295–306 (2005).
21. Huang, W., Cao, X., Biase, F. H., Yu, P. & Zhong, S. Time-variant clustering model for understanding cell fate decisions. Proc. Natl Acad. Sci. USA 111, E4797–E4806 (2014).

22. Marco, E. et al. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proc. Natl Acad. Sci. USA 111, E5643–E5650 (2014).

23. Shin, J. et al. Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis. Cell Stem Cell 17, 360–372 (2015).

24. Ji, Z. & Ji, H. TSCAN: pseudo-time reconstruction and visualization in single-cell RNA-seq analysis. Nucleic Acids Res. 44, e117 (2016).

25. Welch, J. D., Hartemink, A. J. & Prins, J. F. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. Genome Biol. 17, 106 (2016).

26. Haghverdi, L., Buttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods 13, 845–848 (2016).

27. Moignard, V. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat. Biotechnol. 33, 269–276 (2015).

28. Khoramian Tusi, B. & Socolovsky, M. High throughput single-cell fate potential assay of murine hematopoietic progenitors in vitro. Ex. Hematol. https://doi.org/10.1016/exphem.2018.01.005 (2018).

29. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187–1201 (2015).

30. Morrison, S. J. & Weissman, I. L. The long-term repopulating subset of hematopoietic stem cells is deterministic and isolatable by phenotype. Immunity 1, 661–673 (1994).

31. Papayannopoulou, T., Brice, M., Broudy, V. C. & Zsebo, K. M. Isolation of c-kit receptor-expressing cells from bone marrow, peripheral blood, and fetal liver: functional properties and composite antigenic profile. Blood 78, 1403–1412 (1991).

32. Weinreb, C., Wolock, S. & Klein, A. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. Bioinformatics (2017).

33. Weinreb, C., Wolock, S., Khoramian Tusi, B., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single cell snapshots. Proc. Natl Acad. Sci. USA. http://doi.org/10.1073/pnas.1714723105 (2018).

34. Yanez, A. et al. Granulocyte-monocyte progenitors and monocyte-dendritic cell progenitors independently produce functionally distinct monocytes. Immunity 47, 890–902.e4 (2017).

35. Magwene, P. M., Lizardi, P. & Kim, J. Reconstructing the temporal ordering of hematopoietic stem cells is deterministic and isolatable by phenotype. Immunity 1, 661–673 (1994).

36. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381–386 (2014).

37. Bresnick, E. H., Lee, H.-Y., Fujiwara, T., Johnson, K. D. & Keles, S. GATA switches as developmental drivers. J. Biol. Chem. 285, 31087–31093 (2010).

38. Li, P. et al. Regulation of bone marrow hematopoietic stem cell is involved in high-altitude erythrocytosis. Exp. Hematol. 39, 37–46 (2011).

39. Grover, A. et al. Erythropoietin guides multipotent hematopoietic progenitor cells toward an erythroid fate. J. Exp. Med. 211, 181–188 (2014).

40. Mancini, E. et al. FOO-1 and GATA-1 act sequentially to specify definitive megakaryocytic and erythroid progenitors. EMBO J. 31, 351–365 (2012).

41. Koulis, M., Porpiglia, E., Hidalgo, D. & Socolovsky, M. in A Systems Biology Approach to Blood, Vol. 844 (eds Corey, S. J. et al.) Ch. 3, 37–58 (Springer New York, 2014).

42. Agosti, V., Karur, V., Satyanarayana, P., Besmer, P. & Wojcikowski, D. M. A KIT juxtamembrane PY567-directed pathway provides nonredundant signals for erythroid progenitor cell development and stress erythropoiesis. Exp. Hematol. 37, 159–171 (2009).

43. Kouy, M. J. & Bondurant, M. C. Erythropoietin retards DNA breakdown and prevents programmed death in erythroid progenitor cells. Science 248, 378–381 (1990).

44. Yee, K., Bishop, T. R., Mather, C. & Zon, L. I. Isolation of a novel receptor tyrosine kinase cDNA expressed by developing erythroid progenitors. Blood 82, 1335–1343 (1993).

45. van den Akker, E. et al. Tyrosine kinase receptor RON functions downstream of the erythropoietin receptor to induce expansion of erythroid progenitors. Blood 103, 4457–4465 (2004).

46. Pop, R. et al. A key commitment step in erythropoiesis is synchronized with the cell cycle clock through mutual inhibition between PU.1 and S-phase progression. PLoS Biol. 8, e1000484 (2010).

47. Hwang, Y. et al. Global increase in replication fork speed during a p57KIP2-regulated erythroid cell fate switch. Sci. Adv. 3, e1700298 (2017).

48. Shearstone, J. R. et al. Global DNA demethylation during mouse erythropoiesis in vivo. Science 334, 799–802 (2011).

49. Neneth, E. & Ganz, T. Anemia of inflammation. Hematol. Oncol. Clin. North Am. 28, 671–681 (2014).

50. Liang, R. et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell 13, 1977–2000 (2002).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was funded by a Leukemia and Lymphoma Society Scholar award (1728-13) and R01DK100915 and R01099281 (M.S.). A.M.K. is supported by a BW Fund CASI award and an Edward J. Mallinckrodt Foundation Grant, S.L.W. and C.W. are supported by National Institutes of Health (NIH) training grant 5T32GM080177-07.

Author Contributions M.S. and A.M.K. designed the experiments and supervised the project. B.K.T., S.L.W., Y.H., D.H. and R.Z. performed experiments including inDrops (B.K.T., R.Z., S.L.W.), FACS and antibody screening (B.K.T., D.H.), single-cell fate assays and cell cycle analysis (B.K.T.), western blotting (Y.H.), qRT–PCR (B.T.K.), pSTAT3/5 (Y.H., D.H.) and colony assays for novel growth factors (Y.H.). S.L.W. and C.W. performed single-cell data analysis, informatics and PBA (B.T.K.), pSTAT3/5 (Y.H., D.H.) and colony assays for novel growth factors (Y.H.). A.M.K. and S.L.W. are supported by National Institutes of Health (NIH) training grant 5T32GM080177-07.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to A.M.K. (allon_klein@hms.harvard.edu) or M.S. (merav.socolovsky@umnassmed.edu).

Reviewer Information Nature thanks B. Göttgens, F. Hamey and the other anonymous reviewer(s) for their contribution to the peer review of this work.
**METHODS**

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

**Ethical compliance.** All mouse experiments described in this project fully comply with the mouse protocol issued to the Socolovsky laboratory by the Institutional Animal Care and Use Committee (IACUC) of the University of Massachusetts Medical School.

**Mice and RNA-seq.** For the basal bone marrow (bBM) sample, and for the sorted P1–P5 populations, bone marrow was collected from 8-week-old adult BALB/c female mice (Jackson Laboratories). For the EPO-stimulated bone marrow (eBM) sample, 8-week-old adult BALB/c female mice were injected with EPO (Procr, Amgen) subcutaneously once every 24 h for a total of 48 h, at 100 U per 25 g body weight. For the fetal liver (FL) sample, BALB/c female mice were set up for timed pregnancies, and fetal livers were collected on embryonic day 13.5.

**Cell preparation for scRNA-seq.** Tissue collection. For bone marrow preparation, femurs and tibiae were collected immediately following euthanasia, and placed in cold (4 °C) ‘staining buffer’ (PBS containing 0.2% bovine serum albumin (BSA) and 0.08% glucose). Bones were flushed using a 2-ml syringe with a 26-gauge needle and then crushed with a pestle and mortar to obtain all cells. Collected bone marrow cells were filtered through a 40-µm strainer and washed in cold ‘Easy Sep’ buffer (PBS; 2% fetal bovine serum (FBS); 1 mM EDTA). Fetal livers were prepared by mechanical dissociation in staining buffer and a washing in ‘Easy Sep’ buffer.

**Positive selection for Kit+ cells.** Bone marrow and fetal liver cell samples were each enriched for Kit+expressing cells using magnetic beads, with the Mouse Biotin Selection Kit (STEMCELL Technologies, 18556) and Biotin Rat Anti-Mouse CD117 Antibody (clone 2B8, BD Bioscience), following the manufacturer’s protocol.

**Density gradient centrifugation.** Following magnetic bead selection, dead cells and debris were removed from the bone marrow and fetal liver samples using density centrifugation in Optiprep (Sigma, D1556). In brief, cells were re-suspended in 0.5 ml staining buffer, mixed with 1 ml of 40% of Optiprep in PBS, and placed in a 5-ml tube. The cell suspension was carefully over-layered with 2 ml of 20% Optiprep solution, and 1 ml of 5% Optiprep solution, and centrifuged at 800 g for 15 min (centrifuge break off). The top visible cell band that formed during centrifugation contained the live, Kit+ single cells (confirmed by flow cytometric analysis). This layer was carefully aspirated and used directly in the inDrops platform.

**Single-cell transcriptome droplet microfluidic barcoding using inDrops.** For scRNA-seq, we used inDrops following a previously described protocol with the modifications summarized in Supplementary Table 7. Following droplet barcoding reverse transcription, emulsions were split into aliquots of approximately 1,000 single-cell transcriptomes and frozen at −80 °C. Two batches of Kit+ libraries were prepared, referred to as batch 1 (bBM, n = 840 cells; eBM, n = 1,141 cells; FL, n = 1,953 cells) and batch 2 (bBM, n = 4,592 cells; eBM, n = 1,314 cells; FL, n = 7,529 cells) in Supplementary Table 7. These cell numbers correspond to the final number of transcriptomes detected upon sequencing (see ‘Cell filtering and data normalization’), and were in agreement with estimated inputs.

For the FACs subsets P1, P1–CD71high, P2, P3, and P5 (referred to collectively as ‘P1–P5’), all libraries were prepared in parallel, with a total of 16,206 cell barcodes detected in the sequencing data before filtering (P1, n = 5,733 cells; P1–CD71high, n = 1,631 cells; P2, n = 2,630 cells; P3, n = 2,101 cells; P4, n = 1,589 cells; P5, n = 2,522 cells).

**Sequencing and read mapping.** The first batch of Kit+ (bBM, eBM and FL) libraries was sequenced on a HiSeq 2000, the remaining Kit+ libraries were sequenced on three NextSeq 500 runs, and all P1–P5 libraries were sequenced on a single NextSeq 500 run. Raw sequencing data (FASTQ files) were processed using the previously described inDrops.py bioinformatics pipeline (available at https://github.com/indrops/indrops), with a few modifications. Bowtie v.1.1.1 was used for mapping reads (UMIFM) in at least three cells), and which contributed to principal components with eigenvalues greater than those obtained after data randomization (n = 59, n = 35 and n = 71 principal components for bBM, eBM and FL samples, respectively). The expression level for each gene was standardized by a z-score transform (mean-subtraction, scaling by s.d.), followed by density-based clustering (DBSCAN) on a 2D PCA–tSNE plot (principal component analysis (PCA) followed by a distributed stochastic neighbour embedding (tSNE)). The tSNE algorithm perplexity parameter was set to 30. Examination of the expression of marker genes in each cluster was then used to identify putative doublets and contaminating cell types.

In the bBM sample, two doublet clusters were identified: one co-expressed markers of mature macrophages and erythrocytes (n = 38 cells), and the other co-expressed markers of granulocyte and erythroid progenitors (n = 75 cells). The eBM sample included a cluster of mature macrophages (n = 40 cells) but no identifiable cluster of doublets. The FL sample contained four contaminating cell types: vascular endothelium, hepatocytes, mesenchymal cells and mature macrophages (n = 769 cells total), in addition to a small cluster of doublets (n = 18 cells). Doublets and contaminant cells were excluded from downstream analyses.

To increase confidence that putative doublet clusters were indeed combinations of two single cells, rather than true intermediate or transitional states, we generated simulated ‘artificial’ doublets by randomly sampling and combining observed transcriptomes.

We then applied PCA–tSNE clustering as described earlier to the union of observed and simulated cells, and identified clusters that were primarily composed of cells with a large number of doublet neighbours (two clusters in bBM, one in FL). These clusters were the same putative doublet clusters identified in the previous paragraph.

**Batch correction.** Within each Kit+ sample, we observed batch effects between the first and second sequencing runs, with slightly fewer genes detected per cell in the second run compared to the first run. This was consistent with the choice of lower sequencing depth used in the second set of runs, but could also reflect differences in library preparation despite all cells being collected in a single droplet run. To prevent batch effects from distorting subsequent data analysis, for each sample we used the second (larger) batch to select variable genes and to calculate principal component gene loadings. Cells from all batches were then projected into the reduced space, and all subsequent analysis was performed on the reduced principal component space.

**Data visualization and construction of k-nearest neighbour graphs.** After cell filtering, data were prepared for visualization and PBA by constructing a k-nearest neighbour (kNN) graph, in which cells correspond to graph nodes and edges connect cells to their nearest neighbours. A kNN graph was constructed separately for each of the three Kit+ samples and for the merged P1–P5 samples (note that the kNN graph for P1–P5 was used only for the visualization in Extended Data Fig. 3).

For the Kit+ samples, genes with mean expression >0.05 and coefficient of variation >2 were used to perform PCA down to 60 dimensions (bBM, eBM and FL). For all analyses in this paper, data were z-score normalized at the gene level before PCA (qualitatively similar results were also obtained without z-score normalization, which weights highly expressed genes more heavily than lowly expressed genes). After PCA, a kNN graph (k = 5) was constructed by...
connecting each cell to its five nearest neighbours (using Euclidean distance in the principal component space).

For P1–P5, highly variable genes were filtered using the $v$-score statistic (above-Poisson noise) rather than the coefficient of variation, keeping the top 25% most variable genes and requiring at least three UMIFM to be detected in at least three cells ($n = 3,459$ genes). Additionally, a strong cell cycle signature was observed in the initial graph visualization, manifested by the co-localization of cells expressing G2/M genes (Ube2c, Hmg2b, Hmg2a, Tuba1b, Mkia67, Ccl4l1, Top2a, Tubb4b). Therefore, we constructed a G2/M signature score by summing the average $v$-score of these genes, then removed genes that were highly correlated (Pearson $r > 0.2$) with the signature ($n = 31$ genes). Finally, the kNN graph was constructed with $k = 4$ using the first 30 principal components.

The kNN graphs were visualized using a force-directed layout using a custom interactive software interface called SPRING$^{39}$. For the kIt$^{+}$ samples, several manual steps were taken to improve visualization. It is important to emphasize that the manipulations affect visualization only. All subsequent analyses depend on the graph adjacency matrix, which is not affected by any of the changes to the graph layout. For visualization purposes, we manually extended the length of the megakaryocytic, basophlic and monocytic branches by pinning the position of cells at the end of each branch, and allowing the remaining structure to follow. In the bBM sample, we compressed the CEP ‘bulge’ region of the graph by bringing its bounding cells together.

**Smoothing over the kNN graph.** We smoothed data over the kNN graph for gene expression visualization and for one analysis (see ‘Global changes in gene expression in stress conditions’). Smoothing was done by diffusing the property of interest (for example, gene expression counts or number of mapped cells) over the graph, as described$^{35}$. In brief, let $A$ be the adjacency matrix of the kNN graph, in which $A_{ij} = 1$ if an edge in the graph connects nodes $i$ and $j$. Define $A$ as the transition matrix, obtained by row-normalizing $A$:

$$A'_{ij} = \frac{A_{ij}}{\sum A_{ij}}$$

Let $E_i$ be the quantity of interest (for example, expression level) in cell $i$. Then $E', the smoothed vector of $E$, is computed as follows:

$$E' = \gamma (I - (1 - \gamma) A')^{-1} E$$

in which $\gamma$ is a diffusion constant ($\gamma = 0.05$ in all presented analyses) and $I$ is the identity matrix.

**Formal measure of the continuity of transcriptional states.** To demonstrate that the continuous appearance of the kIt$^{+}$ transcriptomes was not a trivial outcome of our analysis methods, we used the same tools to analyse an scRNA-seq data set of mature blood cells (peripheral blood mononuclear cells, PBMCs) (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.0.1/pbmc8k), which consist of several distinct cell types (Extended Data Fig. 1a). In addition to generating a SPRING plot of the data, we also assessed the interconnectivity of each dataset by examining the behaviour of random walks over the kNN graphs, as previously described$^{36}$. In detail, after subsampling the PBMC data to contain the same number of cells as the bBM dataset, we applied PCA and constructed a kNN graph ($k = 10$) for each dataset. We then simulated 1,000 random walks for each graph and plotted the fraction of nodes (cells) visited as a function of the number of steps (Extended Data Fig. 1a).

**PBA.** The PBA algorithm calculates a scalar ‘potential’ for each cell that is analogous to a distance, or pseudotime, from an undifferentiated source, and a vector of fate probabilities that indicate the distance to fate branch points. These fate probabilities and temporal ordering were computed using the Python implementation of PBA (available online from https://github.com/AllonKleinLab/PBA), as described$^{33}$. The inputs to the PBA scripts are a set of comma-separated values (csv) files encoding: the edge list of a kNN graph of the cell transcriptomes (A.csv); a vector assigning a net source/sink rate to each graph node (R.csv); and a lineage-specific binary matrix identifying the subset of graph nodes that reside at the tips of branches (S.csv). These files are provided in the Supplementary Data for the bBM and D datasets. In PBA, the graph is run according to the following steps:

1. Apply the script `compute_linnpyv.py -A A.csv`, here inputting edges (flag ‘-e’) from the SPRING-kNN graph (see above). This step outputs the random-walk graph Laplacian, Linn.py.

2. Apply the script `compute_linnpyv.py -L Linn.py -R R.csv`, here inputting the inverse graph Laplacian (flag ‘-L’) computed in step (1) and the net source/sink rate to each graph node (flag ‘-R’). This step yields a potential vector (Vnpy) that is used for temporal ordering (cells ordered from high to low potential). The vector $R$ provided in the Supplementary Data was estimated as described in the next section.

3. Apply the script `compute_fate_probabilities.py -S S.csv -V Vnpy -e A.csv -D D'`, here inputting the lineage-specific exit rate matrix (flag ‘-S’), the potential (flag ‘-V’), computed in step (2), the same edges (flag ‘-e’) used in step (1) and a diffusion constant (flat ‘-D’) of 1. This step yields fate probabilities for each cell.

Figures 1–6 make use of PBA analyses of bBM data. For Fig. 4e and Extended Data Fig. 8, a temporal ordering of erythroid differentiation was generated for the FL dataset using the same steps, with input files that are also provided in Supplementary Data.

**Estimation of net source/sink rate vector $R$.** A complete definition of the vector $R$ in terms of biophysical quantities has been published previously$^{35}$. In brief, for a gene expression space described by a vector $x = (x_1, x_2, \ldots, x_d)$ giving the expression of each of $d$ genes, $x$ is the net imbalance between cell division and cell loss locally for cells with gene expression profile $x$. $R(x)$ is corrected for cell enrichment and loss resulting from experimental procedures such as sample enrichment, as follows. In this experiment, all progenitors including HSCs express Kit, but eventually downregulate it as they terminally differentiate. Thus, no cells enter the experimental system other than through proliferation of existing Kit$^{+}$ HPCs, but the selection for Kit$^{+}$ cells during sample isolation induces a net sink on cells downregulating Kit expression. For a self-renewing system, cell division and cell loss are precisely balanced, so $R(x) |_{d = 0}$. To apply PBA, one does not need to estimate $R(x)$, but only its value at points $x$ at which the $M$ cells $i = 1,…, M$ are observed in the scRNA-seq measurement. Thus $R$ is a vector over the cells in the system. For a self-renewing system, the sum over all cells satisfies the same constraint, $\sum R_i = 0$.

**Estimation of $R$.** We assigned negative values to $R$ for the ten cells with the highest expression of marker genes for each of the seven terminal lineages (see Supplementary Table 1 for marker genes), which were separately confirmed to show reduced Kit expression. We assigned different exit rates to each of the seven lineages using a fitting procedure that ensured that cells identified as putative HSCs would have a uniform probability to become each fate. Putative HSCs were identified by the similarity of their transcriptomes to microarray profiles from the ImmGen database (we used SC.LT34FBM (long-term bone marrow HSCs) for bBM and SC.STSL.FL (short-term FL HSCs) for FL; for more details, see section ‘ImmGen Bayesian classifier’). We assigned a single positive value to all remaining cells, with the value chosen to enforce the steady-state condition $\sum R_i = 0$. In the fitting procedure, all exit rates are initially set to one and iteratively incremented or decremented until the average fate probabilities of the putative HSCs were within 1% of uniform. The resulting vector $R$ is provided in the Supplementary Data. The separate lineage exit rates were then used to form the lineage-specific exit rate matrix $S$, also provided in the Supplementary Data.

**Assignment of PBA fate probabilities and temporal ordering to eBM dataset.** For each of the eBM cells we assigned the average temporal order (or potential V) and average fate probabilities of the 20 mostly similar bBM cells. To do this, we first carried out a PCA on the bBM cells into 60 dimensions. We then used the gene loadings of the 60 principal components to project the eBM data into the same principal component space. The distance of each eBM cell to each bBM neighbour was then measured by cosine distance in the 60-dimensional sub-space.

**ImmGen Bayesian classifier.** We used a published microarray profile$^{36}$ to search for similar cells in our own dataset using a naive Bayesian classifier, implemented as follows.

The Bayesian classifier assigns cells to microarray profiles on the basis of the likelihood of each microarray profile for each cell, with the likelihood calculated by assuming that individual mRNA molecules in each cell are multinomially sampled with the probability of each gene proportional to the microarray expression value for that gene. Consider a matrix $E$ of microarray counts (UMIs) $n$ rows (for cells) and $g$ columns (for genes), and also a matrix $M$ with $m$ rows (for microarray profiles) and $g$ columns (for genes). $M$ was quantile normalized and then each microarray profile was normalized to sum to one. Normalization of $E$ was performed as described earlier (see ‘Cell filtering and data normalization’). The $n \times m$ matrix $S_p$, giving the likelihood of each microarray profile $p$ for each cell $i$, is

$$S_p = Z_p \prod_{k=1}^{g} M_{ik}^{S_{pk}}$$

where $Z_p$ is a normalization constant that ensures that $\sum S_p = 1$.

**Computing the haematopoietic lineage tree.** We used the fate probabilities from PBA to infer the topology of the haematopoietic lineage tree using an iterative approach (Fig. 1e, f). Each iteration began with a set of fates and a probability distribution over those fates for each cell. For every pair of fates, we computed a fate coupling score (see later) and merged pairs with a score significantly higher than expected under a null model. The merged fates inherited probabilities from the starting fates by simple pairwise addition.
The coupling score between two fates $A$ and $B$ is the number of cells with $P(A)P(B) > \varepsilon$, in which we used a value $\varepsilon = 1/14$ throughout. To generate a null distribution for each fate pair, we computed pairwise coupling scores for 1,000 permutations of the original fate probabilities. The heat maps in Fig. 1e show $z$-scores with respect to these null distributions.

**Analysis of fate-correlated genes at haematopoietic cell choice points.** To discover fate-associated genes at key choice points in haematopoiesis (Extended Data Fig. 2, Supplementary Table 2), we ranked transcription factors and cell-surface markers (transcription factors from http://genome.gsc.riken.jp/TFdb/tf_list.html, cell-surface markers from https://www.thermosisher.com/us/en/home/life-science/cell-analysis/cell-analysis-learning-center/cell-analysis-resource-library/ebioscience-resources/mouse-cd-other-cellular-antigen.html) by their correlation with PBA-predicted fate probability, restricting to cells that were bipotent for the given choice. Specifically, to find transcription factors associated with fate $A$ at an A/B choice point, we first selected cells with $P(A) \times P(B) > \varepsilon$, and then ranked the transcription factors by their correlation with the fate bias $(P(A) - P(B))$. In Supplementary Table 2, we report all genes with Bonferroni-corrected $P < 0.01$ (Pearson correlation coefficient). In Extended Data Fig. 2, we show at most ten genes for any one choice point.

**Mapping P1–P5 subsets to the Kit+ graphs.** For Fig. 2b, cells from subsets P1–P5 were projected into the same principal component space as the bBM data, then mapped to their most similar Kit+ neighbours. In detail, counts were first converted to TPM for all samples. Then, using only the bBM cells, the 3,000 most variable genes (measured by $t$-score) with at least three UMIFM in at least three cells were $z$-score normalized and used to find the top 50 principal components. Next, the P1–P5 subset cells were $z$-score normalized using the gene expression means and s.d. from the bBM data and transformed into the bBM principal component space. Lastly, each P1–P5 cell was mapped to its closest bBM neighbour in principal component space (Euclidean distance).

**Extracting MPP-to-erythroid trajectory cells.** To isolate the erythroid trajectory, we defined an MPP-to-erythroid axis in each of the three Kit+ datasets by ordering cells on the basis of their graph distance from unbiased MPPs (cells identified on the basis of the ImmGen classifier as described earlier), and keeping only cells for which the probability of erythroid fate increased or remained constant with graph distance. Graph distance was measured by PBA potential, and starting with the cell closest to the HSC origin, we added the cell with next highest potential to the trajectory if the PBA-predicted erythroid probability for cell $i$ was at least 95% of the average erythroid probability of the cell(s) already in the trajectory.

More formally the procedure is as follows: order all $N$ cells in the experiment from highest to lowest PBA potential $V$, with decreasing potential corresponding to increasing distance from MPPs. Let $E_i$ be an indicator variable for the membership of ordered cell $i$ in the erythroid trajectory ($E_i = 1$ if cell $i$ is in the trajectory; otherwise, $E_i = 0$). If $P_i$ is the PBA-predicted erythroid probability for ordered cell $i$, then $E_i = 1$ if

$$P_i > 0.95 \times \frac{\sum_{k<i} P_k \times E_k}{\sum_{k<i} E_k}$$

Cells on the erythroid trajectory were then ordered by decreasing potential. Defining $t_j$ as the index of the jth erythroid-trajectory cell,

$$t_j = 1 + \sum_{k<j} E_k$$

Throughout this Article, we report this cell order (akin to the ‘pseudotime’ in other publications) as a percentage of ordered cells, with the first, least differentiated cell at 0% and the most mature cell at 100%. This is not meant to suggest that erythroid differentiation ends with this final observed cell.

**Identifying dynamically varying genes.** For each gene, a sliding window ($n = 100$ cells) across the MPP-to-erythroid ordering was used to identify the windows with maximum and minimum average expression as previously described. A t-test was then performed to assess the statistical significance of the difference in expression levels. To estimate the false discovery rate (FDR), we permuted the order of the cells and repeated the above analysis. For a $P$ value generated by the observed (non-permuted) ordering, the FDR-corrected $P$ value is the fraction of genes from the permuted ordering with that $P$ value or less. Any gene with an FDR-corrected $P < 0.05$ was considered significantly variable.

**Identifying stage transitions in the MPP-to-erythroid trajectory.** Transition points between stages of erythropoiesis were defined using the frequency of gene inflection points (Fig. 4b), patterns of PBA-predicted fate probabilities (Fig. 1c), and the fate potentials of FACS subsets P1–P5 (Figs 2, 3). However, owing to the continuous nature of the transcriptional states, the locations of these transitions should be considered approximate.

The inflection point density is the number of genes turning on or off at a given point on the trajectory. For each gene, inflection points were identified as the points with maximally increasing or decreasing expression as follows. First, the trajectory of each dynamically varying gene was smoothed using Gaussian smoothing with a width $\sigma = 5\%$ of total trajectory. The gene expression derivative for gene $k$, denoted $x_k$, was then computed by taking a ten-cell moving average of the difference between consecutive smoothed gene expression values. Inflection points were then identified as the points with maximum or minimum derivatives for each gene. To exclude maxima resulting from relatively small fluctuations in gene expression, only appreciably large extrema were kept for further analysis. Specifically, the point with the maximum derivative for gene $k$, max($x_k$), was kept only if

$$\max(x_k') > \text{median}(\text{abs}(x_k'))$$

Minima were similarly filtered, requiring the ratio to be $< -Q$. We chose a threshold $Q = 6$, but results do not qualitatively change over a range of $Q$. We then plotted the density of these inflection points over the MPP-to-erythroid axis. Regions with large-scale changes in gene expression have a high density of inflection points, whereas a low density characterizes relatively stable states.

**Dynamic gene clustering.** Dynamically varying genes were clustered on the basis of their behaviour at the transition points. To prevent overfitting, we used only three transitions (3%, 18% and 86%) by splitting the EEP state and assigning the first and second halves to the EBMP and CEP states, respectively. At each transition, genes were classified as increasing, decreasing or unchanging, giving a total of $3^3 = 27$ possible patterns. After smoothing gene expression traces, the data were binned by calculating the mean expression in each of the four stages. To remove noisy genes or genes that varied little across bins, we calculated the range of binned expression values, range($x_{bin}$) = max($x_{bin}$) − min($x_{bin}$), for each gene and proceeded with the top 50% most variable genes. Next, to place all genes on a similar scale, the binned expression values of each gene were divided by the maximum binned value of that gene. Finally, the differences between consecutive bins were thresholded: differences that were greater than 0.15 were called increasing, differences that were less than −0.15 were called decreasing and differences that were between −0.15 and 0.15 were called unchanging.

**Gene set enrichment analysis.** Each of the 27 gene clusters was used as input for gene set enrichment analysis (GSEA) (hypergeometric test), using all genes as background. Ribosomal genes were excluded from the input, as were predicted genes (gene names starting with ‘Gm’). Gene sets from the following lists of the MSigDB v5.1 (ref. 61) dataset were tested for enrichment: Hallmark (h.all.v5.1symbols.gmt), C2 curated canonical pathways (c2.cp.v5.1.symbols.gmt), C3 transcription factor targets (c3.tft.v5.1.symbols.gmt), and C5 Gene Ontology (c5.all.v5.1.symbols.gmt). Additionally, for the transcription factor target enrichment analysis, we used gene sets from the ChEA database.

**Cell cycle phase analysis.** Genes with periodic expression correlated with the cell cycle in HeLa cells were used to generate a cell cycle phase score for each cell. The list of phase-specific genes was filtered to exclude genes with a mean expression $> 25$ TPM in cells on the MPP-to-erythroid trajectory. For Fig. 6a, a sliding window average was computed using a window size of 10% MPP-to-erythroid progression (~200 cells) and a jump size of 5%. For Fig. 6b, counts were normalized by the mean expression at the gene level, and smoothed using a Gaussian kernel. Then, a phase score was calculated for each phase (G1/S, S, G2/M, M/G1) by averaging the smoothed gene expression traces for the genes specific to that phase.

**Testing the influence of cell cycle genes on the MPP-to-erythroid cell order.** To test the extent to which cell cycle genes influenced the ordering of cells along the MPP-to-erythroid trajectory, we excluded annotated cell cycle genes (described in Table 2) from the gene ontology database (GO:0007049) and Cyclebase, and repeated kNN graph construction and PBA. As shown in Extended Data Fig. 10, the resulting cell order was largely unchanged, as were the dynamics of cell cycle genes.

**Identifying genes that change steadily in the CEP stage.** To identify genes that are steadily up- or downregulated throughout the CEP (Fig. 6c, Supplementary Table 6), we tested the magnitude of change (slope) and the linearity of change (the error of the actual gene trace from a straight line) for each gene. Restricting analysis to cells in the CEP stage and genes with at least two UMIFM in at least five cells, we fit a linear regression to the ordered gene expression values and also generated a smoothed expression trace using a Gaussian kernel (width $\sigma = 5\%$). We then computed a ‘linearity score’ for each gene by dividing the slope of the regression line by the root-mean-square error between the regression line and the smoothed trace. Steadily increasing genes receive large positive scores, whereas

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
steadily decreasing genes are assigned large negative scores. Genes that do not change much or that change non-linearly (for example, sharply increasing only at the end of the stage) receive scores close to 0.

**Global changes in gene expression in stress conditions.** Cells from eBM and FL (stress samples) were mapped to their most similar BBM counterparts, and differentially expressed genes were identified. Mapping was carried out by applying PCA to the BBM and stress samples and finding the closest 20 BBM neighbours for each stress cell. Specifically, the input genes were the principal variable genes described in the 'Cell filtering and data normalization' section. Count matrices were z-score normalized separately for each sample, and PCA was performed on the basal sample to obtain the gene loadings. Using the top 60 principal components, each sample was then transformed using these coefficients, thereby projecting the cells into the same PCA space. To validate this mapping method, we performed the same procedure using different subsets of BBM data as training and test sets (see 'Validation of cross-sample cell mapping' section).

The 20 closest BBM neighbours (Euclidean distance) of each stress cell were found, and for the purpose of comparing gene expression, each of these k (20) neighbours inherited 1/k (1/20) of the transcript counts from the mapped stress cell. To enable the comparison of regions of gene expression space (as opposed to comparing single mapped cells to single basal cells), the mapped and original gene expression values were smoothed over the kNN graph, as described in 'Smoothing over the kNN graph'. To avoid comparing gene expression patterns in regions that were relatively unpopulated in the stress sample (for example, parts of the granulocyte-erythroid lineage), we smoothed the number of mapped stress cells per basal cell over the graph and then excluded basal cells with few mapped stress cells (number mapped cells ≤ 9 for eBM and ≤ 20 for FL).

A differential expression score for each cell i and gene j was defined as the maximum-normalized difference between mapped and basal expression, $S_{ij}^m - S_{ij}$, respectively:

$$d_{ij} = 0.5 \times (\max(S_{ij}^m) + \max(S_{ij}))$$

A gene level score, $D_j = \sum_i d_{ij}$, was computed by summing over the cells. $D_j > 2 \times \sigma_D$ or $D_j < -2 \times \sigma_D$, in which $\sigma_D$ is the average over all gene level scores $D_j$ and $\sigma_D$ is the standard deviation. Then, for each differentially expressed gene, the gene was counted as differentially expressed at a given cell if $d_{ij} > 0.5 \times \sigma_D$, or $d_{ij} < -0.5 \times \sigma_D$, in which $\sigma_D$ is the 99th percentile of $D_j$ and $\sigma_D$ is the 1st percentile of $D_j$.

**Validation of cross-sample cell mapping.** To test the accuracy of the method for mapping eBM and FL cells to BBM cells, we divided the BBM sample into a training set (random sample of 75% of the cells) and test set (the remaining 25%). We then computed the gene loadings. Using the top 60 principal components, $x^m_{ij}$ was defined as the average over all gene level scores $D_j$ and $\sigma_D$ is the standard deviation. Then, for each differentially expressed gene, the gene was counted as differentially expressed at a given cell if $d_{ij} > 0.5 \times \sigma_D$, or $d_{ij} < -0.5 \times \sigma_D$, in which $\sigma_D$ is the 99th percentile of $D_j$ and $\sigma_D$ is the 1st percentile of $D_j$. Intuitively, this score is large and positive if a gene is more strongly upregulated within the ROI than without, large and negative if a gene is more strongly downregulated within the ROI than without and close to 0 otherwise.

**To build gene lists for GSEA input, we first selected genes with score, $> 0.1 \times \max(score)$ (for upregulated) or score, $> 0.1 \times \max(score)$ (for downregulated) and then used the top 100 genes by binomial test P value. Flow cytometric sorting for P1–P5 subsets.** A detailed protocol of this procedure can be found at the Protocol Exchange.

Bone marrow cells from adult BALB/c male or female mice (aged 8–12 weeks) were lineage-depleted using the Mouse Streptavidin RapidSpheres Isolation Kit (STEMCELL Technologies 19860A), with the following biotinylated antibodies: anti-CD11b (clone M1/70, BD Biosciences 553782), anti-CD11c (clone 1D3, BD Biosciences 553784), anti-TER119 (clone TER119, BD Biosciences 553125), anti-CD4 (clone RM4-5, BD Biosciences 553045), anti-CD8A (Ly-2) (clone 53-6.7, BD Biosciences 553029), anti-CD19 (clone CD1D, BD Biosciences 553784), anti-TER119 (clone CD11b, BD Biosciences 553672).

**Data analysis.** Lineage-depleted cells were then labelled with the following antibodies in the presence of 1% rat serum: streptavidin Alexa Fluor 488 (Molecular Probes) to mark lineage-positive cells, CD117–APC Cy7 (clone 2B8, Biogenoid 108526), TER119–BUV395 (clone TER-119, BD Biosciences 563827), CD71–PE Cy7 (clone R7217, Biogenoid 113812), CD55–AF647 (clone RIKO-3, Biogenoid 131806), CD105–PE (clone M77/18, Biogenoid 120408), CD150–BV650 (clone TC15-12F12.2, Biogenoid 115931), CD41–BV605 (clone MWRG30, Biogenoid 133921), CD49f (also known as ITGA6)–BV421 (clone GoH3, Biogenoid 313624).

After washing, cells were re-suspended in DAPI-containing buffer and sorting was performed using a BD FACSAria II with a 100-μm nozzle. Sorted populations were defined as in Fig. 2a.

**qRT–PCR on sorted populations.** RNA was prepared from sorted cell subsets using the RNeasy Micro Kit (Qiagen 74004) or TRizol reagent (Ambion 15956026), and measured with RiboGreen RNA reagent kit (Thermo Fisher Scientific) on the 3300 NanoDrop Fluorospectrometer. cDNA was synthesized using the same amount of input RNA for all samples in a parallel reaction, using the Super Script III first-strand synthesis system for RT–PCR (Invitrogen) with random hexamer primers. The ABI 7300 sequence detection system, TaqMan reagents and TaqMan MGB probes (Applied Biosystems) were used following the manufacturer’s instructions. Quantitative PCR was carried out on four serial dilutions of each CDNA sample, and the linear part of the template dilution/signal response curve was used to calculate relative mRNA concentrations following normalization to Actb3, using the ΔΔCt method.

The following TaqMan MGB probes were used: Mst1r (Mm00436382_m1), Ryk (Mm01238551_m1), Il17ra (Mm00432414_m1), Mif2 (Mm00809556_s1), Slc26a1 (Mm01198850_m1), Slc4a1 (Mm00414492_m1), Tnib2 (Mm0045876_m1), Cd34 (Mm00519283_m1), Mesi1 (Mm00487664_m1), Hpo (Mm01152654_m1), Pf4

After mapping stress cells to their single closest neighbour in the BBM sample (as described in the previous section), we selected BBM cells in the ROI and the stress cells mapping to them. We first identified genes differentially expressed within the ROI by performing a binomial test for differential expression, which tests the probability that a gene is expressed more frequently in one population than another. After correcting for multiple hypothesis testing (Benjamini–Hochberg procedure), we proceeded with genes with an FDR-corrected P < 0.05.

To identify genes differentially expressed specifically within the ROI and not elsewhere, we calculated the mean-normalized expression difference for ROI cells and non-ROI cells for the genes found to be significant in the binomial test. For two samples, A (stress) and B (basal), the mean-normalized expression difference of gene i within the ROI, $y_{in}$ is

$$y_{in} = \frac{x_{in} - x_{in}^B}{\sigma_{x_{in}} + \sigma_{x_{in}^B}}/2$$

in which $x_{in}$ is the average expression of gene i within the ROI in sample A. A similar score was calculated for cells outside the ROI:

$$y_{out} = \frac{x_{out} - x_{out}^B}{\sigma_{x_{out}} + \sigma_{x_{out}^B}}/2$$

Plotting $y_{in}$ against $y_{out}$ clearly reveals genes that are more highly differentially expressed within the ROI than without. A single score per gene was computed as follows:

$$score = \begin{cases} \max(y_{in} - y_{out}, 0) & \text{if } y_{in} > 0 \\ \min(y_{in} - y_{out}, 0) & \text{if } y_{in} < 0 \end{cases}$$

Intuitively, this score is large and positive if a gene is more strongly upregulated within the ROI than without, large and negative if a gene is more strongly downregulated within the ROI than without and close to 0 otherwise.
To generate the Il17ra−/− mice, Il17ra<sup>Δ<sup>lox</sup></sup>-/ mice<sup>ab</sup> were bred with CMV-Cre mice (Jackson Laboratory 03465). The generation of the Il17ra<sup>−/−</sup> allele in the F1 generation of Il17ra<sup>Δ<sup>lox</sup></sup>-/ and CMV-Cre mating pairs was screened by PCR of tail DNA. To remove the CMV-cre allele present in the F1 generation, Il17ra<sup>−/−</sup> CMV-Cre<sup>−/−</sup> mice were outbred with B6 mice.

**Colony-formation assays with human bone marrow**. Human bone marrow from healthy donors (85,000 cells per assay; see Human cells section) were plated in MethoCult (M3234, STEMCELL Technologies) supplemented with EPO (1 U ml<sup>−1</sup>) and IL-3 (50 ng ml<sup>−1</sup>). Colonies were scored after 7 days.

**Flow cytometric cell cycle analysis of bone marrow cells in vivo**. Flow cytometric analyses were carried out as described<sup>43</sup>. In brief, BrdU (100 μl of 10 mg ml<sup>−1</sup> stock in PBS) was injected intraperitoneally into adult mice 30 min before euthanasia. After collection of bone marrow cells, cells were immediately placed in cold staining buffer, labelled using a LIVE/DEAD kit (Invitrogen) to identify dead cells and were then fixed and permeabilized. Cell-surface staining for each of the five subsets P1–P5 was carried out as described earlier. Simultaneously, incorporated BrdU was detected using a biotin-conjugated anti-BrdU antibody (Abcam) following mild digestion with DNAsel. DNA content was assayed by labelling with the fluorescent indicator 7-AAD (BD Biosciences). Cells were then analysed for cell-surface labelling, BrdU incorporation and DNA content by flow cytometry.

**Western blot analysis**. Bone marrow cells were collected and immediately enriched for Kit<sup>−/−</sup> Lin<sup>−</sup> TER119<sup>−</sup> CD71<sup>−</sup> cells using magnetic beads, as described earlier. The enriched cell fraction was initially placed in culture in IMDM with 20% FCS and EPO (2 U ml<sup>−1</sup>), in the presence or absence of IL-17A (R&D Systems 7955–IL–025). For negative controls, we used 3T3-L1 cells. For positive controls, 3T3-L1 cells were collected on days 7, 21, 28 and 35.

**Intracellular signalling by STAT3 and STAT5**. Freshly collected bone marrow cells were enriched for Lin<sup>−</sup> TER119<sup>−</sup> cells using magnetic beads, as described above. The enriched cells were incubated in cytokine-free, low-sodium medium (IMDM with 2% FCS) for 3 h. EPO (0.5 U ml<sup>−1</sup>), IL-7A (20 ng ml<sup>−1</sup>) or both together was then added to the medium for either 30 or 60 min. Cells were collected, washed with PhosphoWash Buffer<sup>46</sup>, stained with a LIVE/DEAD kit (Invitrogen), fixed and permeabilized with Cytofix/Cytoperm Buffer (BD Biosciences), supplemented with 1 mM sodium orthovanadate (Sigma 450243–10G), 1 mM β-glycerophosphate (Sigma G9422–10G) and 1 μM g-1 ml<sup>−1</sup> Microcystin (EMD Millipore 475815–500UG), and Perm/Wash Buffer I.
(BD Biosciences 557885), and frozen in freezing medium (90% FCS, 10% DMSO, 1 mM sodium orthovanadate, 1 mM L-glutamine and 1 μM 1-methyl-1-cysteine). When thawed, cells were re-fixed and permeabilized, incubated with 5% milk and 200 μg mL−1 rabbit IgG (modified from ref. 69), and stained with p-STAT3-AF488 (B-7) (Santa Cruz sc-8059 AF488), p-STAT3-AF647 (pY694) (BD Bioscience 612599), CD71–PE/Cy7 (Biolegend 113812), CD55–PE (Biolegend 131804), CD105–Pacific Blue (Biolegend 120412), CD150–BV605 (Biolegend 115931), CD49f–PE/Dazzle 594 (Biolegend 313626), CD41–BV650 (Biolegend 133921), CD117 (Kit)–APC/H7 (BD Bioscience 560185), strepavidin–AF700 (Invitrogen S21383) and DAPI. Analysis was performed on an LSRII FACS analyser.

**Code availability.** Python scripts are described in the PBA section, and Supplementary Data 1 contains the input data files and code for running PBA on the bone marrow and fetal liver datasets. Code is available at https://github.com/AllonKleinLab/PBA.

**Data availability.** Sequence data that supports the findings of this study have been deposited in the Gene Expression Omnibus (GEO) with the accession code GSE89754. An interactive tool for the interpretation of these data is available at indrops.indrops, https://github.com/AllonKleinLab/SPRING and https://github.com/AllonKleinLab/PBA.

53. Zilonis, R. et al. Single-cell barcoding and sequencing using droplet microfluidics. Nat. Protocols 12, 44–73 (2017).
54. Ester, M., Kriegel, H., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. 2nd International Conference on Knowledge Discovery and Data Mining (Eds Simoudis, E. et al.) 226–231 (AAAI, 1996).
55. Daszykowski, M., Walczak, B. & Massart, D. L. Looking for natural patterns in data: Part 1. Density-based approach. Chemomtr. Intell. Lab. Syst. 56, 83–92 (2001).
56. van der Maaten, L. Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. 15, 3221–3245 (2014).
57. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214 (2015).
58. Weinreb, C., Wolock, S. & Klein, A. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. Bioinformatics https://doi.org/10.1093/bioinformatics/btx792 (2017).
59. Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for detecting significantly mutated pathways in cancer. J. Comput. Biol. 18, 507–522 (2011).
60. Heng, T. S. et al. The Immunological Genome Project: networks of gene expression in immune cells. Nat. Immunol. 9, 1091–1094 (2008).
61. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl Acad. Sci. USA 102, 15545–15550 (2005).
62. Lachmann, A. et al. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics 26, 2438–2444 (2010).
63. Scialdone, A. et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods 85, 54–61 (2015).
64. Santos, A., Wernersson, R. & Jensen, L. J. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. Nucleic Acids Res. 43, D1140–D1144 (2015).
65. Shekhar, K. et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell 166, 1308–1323.e1330 (2016).
66. Benjamin, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. 57, 289–300 (1995).
67. Tusi, B. K. & Socolovsky, M. Novel FACS strategy for identification of early hematopoietic progenitors including BFU-e. CFU-e and erythroid-biased MPPs Protoc. Exch. http://doi.org/10.1038/protex.2018.031 (2018).
68. El Malki, K. et al. An alternative pathway of imiquimod-induced psoriasis-like skin inflammation in the absence of interleukin-17 receptor a signaling. J. Invest. Dermatol. 133, 441–451 (2013).
69. Porpiglia, E., Hidalgo, D., Koulis, M., Tzafriri, A. R. & Socolovsky, M. Stat5 signaling specifies basal versus stress erythropoietic responses through distinct binary and graded dynamic modalities. PLoS Biol. 10, e1001383 (2012).
Extended Data Figure 1 | See next page for caption.
Extended Data Figure 1 | scRNA-seq of Kit+ haematopoietic progenitors for prediction of the early haematopoietic hierarchy.

a, Top, SPRING plot of 7,959 human peripheral blood mononuclear cells (PBMCs) from 10X Genomics (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.0.1/pbmc8k). Clusters were generated by performing spectral clustering on the underlying kNN graph and annotated on the basis of marker genes. NK, natural killer. Random walks over kNN graphs for the PBMC (middle) and Kit+ bone marrow (bottom) datasets. Each plot shows the fraction of nodes (cells) visited for 1,000 simulated random walks. 
b, Top, SPRING plot of 2,855 Lin−Kit+ SCA1− mouse HPCs from a previously published dataset12. Bottom, SPRING plot of 1,656 cells from three mouse haematopoietic progenitor populations (Lin−Kit+ Sca1+, Lin−Kit+ SCA1−, and Lin− Kit+ SCA1− FLK2−CD34+) from a previously published dataset12. Coloured (non-grey) cells indicate expression of lineage-specific genes (see Supplementary Table 7).

c, Schematic of the population balance law, which relates the dynamic velocities of cells to the distribution of states they are in at a moment in time. The law states that in steady state, after accounting for cell division and loss, the flux of cells entering any region of gene expression space equals the flux out of that region. d, Flow diagram of the inputs and outputs of the PBA algorithm. The population balance law is applied to inputs that include single-cell expression data and estimates of cell proliferation and loss rates at each point in gene expression space; inferred outputs include cell dynamics such as fate probabilities and pseudo-temporal ordering.

e, SPRING plot of bone marrow Kit+ cells (Fig. 1) constructed using only the PBA-predicted fate probabilities and differentiation ordering as inputs (n = 4,763 cells from one inDrops experiment). Coloured cells indicate expression of lineage-specific genes as in Fig. 1b. f, SPRING plot of bone marrow Kit+ cells (Fig. 1), with cells coloured by library preparation batch.
Extended Data Figure 2 | Predicting key regulators at haematopoietic choice points. Candidate regulators of fate choice, identified by ranking transcription factors and transmembrane receptors by their correlation with PBA-predicted fate probabilities at key choice points in haematopoiesis. Top-ranked genes are shown; these include many canonical regulators. Candidate genes that have not been previously reported are marked with asterisks. Several candidates participate in more than one fate choice. Insets show SPRING plots coloured by expression of representative genes. Fog1 is also known as Zfp1, B-myb is also known as Mybl2, PU.1 is also known as Spi1 and Oct2 is also known as Slc22a2.
Extended Data Figure 3 | Mapping HPC subsets P1–P5 to the Kit+ SPRING plot using qRT–PCR and scRNA-seq. a, Subpopulations P1–P5 map onto specific regions of the SPRING plot. SPRING plot heat maps for a panel of marker genes (left) and corresponding measured expression for each of the marker genes by qRT–PCR (middle), performed on sorted cell subsets P1–P5, and on EryA (cells undergoing ETD2). Probable mapping of each of the P1–P5 subpopulations on the basis of qRT–PCR is shown on the SPRING plot (right). Bars represent the mean of two independent experiments (circles, triangles or squares). Expression is shown normalized to Actb mRNA. b, c, SPRING plot of single-cell transcriptomes from freshly sorted P1–P5 subsets (Fig. 2a, b). Cells are coloured on the basis of sorted subpopulation (b) or the expression of lineage-specific marker genes (c) (Supplementary Table 7). d, e, SPRING plots of P1–P5 subpopulation cells, coloured on the basis of expression of basophil (d) and mast cell (e) marker genes. The larger number of cells in the P3 region of the graph resolves a split between the two lineages that was not observable in the original Kit+ dataset.
Extended Data Figure 4 | Validation of PBA predictions. a, Megakaryocytic colonies from sorted subsets P1–P5 and from Kit+ CD55− cells, stained for the megakaryocytic marker acetylcholinesterase. Duplicate cultures are shown; representative of two independent experiments. b, Representative flow cytometry plots to assay fate output of single cells in liquid culture (see Fig. 3, including Fig. 3a for experimental design). Each row corresponds to a single clone, with the left column indicating the source subset (P1–P5, CD55−) of the clone and the cell type(s) produced, as inferred from the FACS plots in the remaining columns. These data are representative of 1,158 single-cell clonal cultures, pooled from three independent sorting experiments (complete dataset is shown in Fig. 3b). c, Bulk liquid cultures of freshly sorted P1–P5 subsets and Kit+ CD55− cells in the presence of EPO and a mixture of cytokines supporting myeloid progenitors. On the indicated days, cells were labelled with antibodies as indicated and analysed by flow cytometry.
Extended Data Figure 5 | See next page for caption.
Extended Data Figure 5 | The early erythroid trajectory. a, qRT–PCR for expression of established erythroid regulators in sorted P1–P5 subsets. Expression of each gene is normalized to Actb. Bars represent the mean of two independent experiments (circles). b, Left, representative western blot of GATA1 expression in sorted P1 (subdivided into CD71^medium and CD71^high subsets), P2 and EryA (CD71^+TER119^+FSC^high cells, representative of ETD) cells. 3T3-GATA1, positive control 3T3 cells virally transduced with a GATA1 expression vector; untransduced 3T3 cells were used as negative control. Right, quantification of GATA1 expression (mean) by densitometry. Data points are from two independent western blots. For gel source data, see Supplementary Fig. 1. c, Density of FACS subsets P1–P5 along the erythroid trajectory. Single-cell transcriptomes from each subset were mapped to their most similar counterparts in the Kit^+ data (Fig. 2a, b; cell numbers analysed for each subset are indicated in Fig. 2b). Shown here is the fraction of mapped cells after smoothing with a Gaussian kernel. Also included are CD71^high P1 cells, constituting cells with the 30% highest CD71 expression in that subset (n = 752 cells post-filter). d, Distribution of CD71 expression in P1 (top) and P2 (bottom) cells immediately after sorting (grey) and after 24 h of in vitro differentiation (lavender). Data are representative of two independent experiments. e, Dynamically varying genes along the MPP-to-erythroid axis were clustered on the basis of their behaviour across three transition points. At each transition, gene expression is increased, decreased or unchanged, giving a total of 27 potential dynamic patterns across all three transitions, shown in red. The number of genes corresponding to each dynamic pattern is noted, and z-score-normalized expression traces for each individual gene are shown in black. Selected clusters were further analysed in Extended Data Fig. 6 and are marked with an asterisk and a representative gene.
Extended Data Figure 6 | See next page for caption.
Extended Data Figure 6 | Gene set enrichment on dynamic gene clusters in early erythroid differentiation. a–d. Nine key dynamic gene clusters along the MPP-to-erythroid progression (Extended Data Fig. 5e) are analysed further for Gene Ontology terms. Each cluster is identified by its dynamic pattern with a cartoon of nodes and edges. Each node represents a progenitor stage (in order, MPP, EBMP/EEP, CEP and ETD), connected to the next stage by an edge that either goes up (for increased expression) or down (for decreased expression) (see Extended Data Fig. 5e). a, Number and identity of transcription factors (TFs) whose targets are enriched in the dynamic clusters, as predicted by chromatin immunoprecipitation methods (ChIP-X) experiments62. b, Significance of enrichment for signalling pathways and Gene Ontology gene sets in the dynamic gene clusters (hypergeometric test with Benjamini–Hochberg correction for multiple hypothesis testing). Rep., representative. c, d, Enrichment of transcription factor targets. c, Heat map (−log_{10} of P value of hypergeometric test with Benjamini–Hochberg correction for multiple hypothesis testing) of target gene enrichment for transcription factors (rows) with targets significantly enriched (P < 0.05) in at least one of the nine dynamic clusters (columns, labelled on top) highlighted in Extended Data Fig. 5e. Note that the transcription factor targets shown are based on previous ChIP-X experiments62 and it is possible that unappreciated transcription factor targets occur in early erythropoiesis. d, Gene expression traces over the erythroid trajectory for the transcription factors from c. Rows match those in c.
Extended Data Figure 7 | Quantification of absolute Kit+ cell number in bone marrow after in vivo administration of EPO. a, b, Eight-week-old female mice were injected with either EPO (100 U per 25 g body weight) or saline (basal), once per day for two days. Bone marrow was collected at 48 h. Viable (trypan blue-negative) cells were counted using a TC20 automated cell counter (BIORAD) and stained for Kit, TER119 and CD71 and lineage markers. Data are from two independent experiments, with five mice analysed individually for each group (basal or EPO) in each experiment. a, Representative flow cytometric analysis of either basal or EPO-stimulated bone marrow, gating on Kit+Lin− cells (left) or on proerythroblasts (ProE) and TER119high cells (right; ProE and TER119high cells are sequential stages of ETD). b, Data summary (mean ± s.d.) for all mice (ten in each group). Top, the fraction of all bone marrow cells for each of the flow cytometric gates defined in a. Bottom, the absolute cell count in adult bone marrow for subsets defined in each flow cytometric gate, or for the total number of bone marrow cells. Significant (P < 0.05) P values are shown (two-tailed t-test, unequal variances).
Extended Data Figure 8 | Identification of stage-specific differential gene expression during the erythroid stress response. **a**, Identification of genes that are differentially expressed in EEP cells of either EPO-stimulated bone marrow (eBM, left) or fetal liver (FL, right) samples, compared with basal bone marrow (bBM). One single-cell inDrop experiment per condition. P values were calculated using a binomial test for differential expression (see Methods) and measure the significance of the expression difference. The specific enrichment score (see Methods) measures the degree to which the differential expression is specific to this region of interest (EEPs): positive scores correspond to region-specific upregulation, and negative to region-specific downregulation. Selected genes are highlighted. **b**, The analysis in **a** applied to the CEP stage. One single-cell inDrop experiment per condition. **c**, Stage-specific differential gene expression during stress, comparing EPO-stimulated and fetal liver samples. The heat map shows the number of differentially expressed genes at each stage that show similar or different patterns of upregulation and downregulation in fetal liver and EPO-stimulated bone marrow samples. Representative gene traces are shown on the right.
Extended Data Figure 9 | Localized gene expression and functional response of the erythroid lineage to stimulation of MST1, RYK and IL-17RA. a, b, Predicted expression pattern (a) and confirmation by qRT-PCR (b) for Mst1r, Ryk and Il17ra in basal bone marrow. In a, traces show the smoothed scRNA-seq gene expression of cells from the basal bone marrow (bBM), fetal liver (FL) and EPO-stimulated bone marrow (eBM) samples, arranged along the erythroid trajectory. qRT–PCR data represent the mean (bars) of two independent experiments (circles). c, Complete results for CFU-e and BFU-e colony-formation assays in methylcellulose, supporting the data shown in Fig. 5. Curves show colony numbers in the presence of increasing concentrations of either EPO, or EPO with a ligand (MSP, WNT5A or IL-17A). Error bars show s.d. of two independent experiments, with four replicates per experiment. Where appropriate, data were fitted to a dose–response curve with a Hill coefficient of one. d, Western blot analysis shows that IL-17RA expression peaks in EEP P2 cells, dropping in CEP P1 cells and in the granulocytic branch (which contributes most of the CD55- cells), consistent with the SPRING plots in Fig. 5a. The western blot is representative of two independent experiments. For gel source images, see Supplementary Fig. 1.
Extended Data Figure 10 | Cell ordering independently of cell cycle genes, and evidence of an S-phase-dependent CEP-to-ETD transition in bone marrow erythropoiesis. a, The computational ordering of cells from MPP to ETD is not sensitive to the inclusion or exclusion of annotated cell cycle genes (cell ordering correlation is R = 0.97). b, c, Reproduction of Fig. 6b, c after the exclusion of cell cycle genes shows that the computationally inferred expression dynamics of cell cycle genes during EEP to CEP differentiation are not sensitive to the inclusion or exclusion of annotated cell cycle genes when ordering cells. d–f, Activation of ETD is dependent on the S phase. d, Schematic of experiments shown in e and f that test the link between S-phase progression and the CEP-to-ETD transition. Kit+Lin−CD71− bone marrow cells were cultured in the presence of EPO for 28 h, with or without the presence of the DNA polymerase inhibitor aphidicolin (Aphi). e, Kit+Lin−CD71− bone marrow cells require the S phase to upregulate CD71, an early event in ETD. Left, CD71 high cells fail to appear in the first 10 h if cells are exposed to aphidicolin; they appear as soon as aphidicolin is removed from the medium. Right, cell cycle analysis of the same cells shows that aphidicolin prevented S-phase progression; aphidicolin removal was followed by a full recovery of S-phase progression, with a high fraction of CD71high cells in S phase. Data are representative of three independent experiments. f, Aphidicolin exposure for 10 h delays induction of β-globin (Hbb-b1) by 10 h. Data are representative of two independent experiments. g, CD71 expression (top), cell cycle phase distribution (middle), and intra-S-phase DNA synthesis rate (bottom), for consecutive FACS gates of increasing CD71 in the early stages of erythropoiesis from the fetal liver (left, representative of four independent experiments) and EPO-simulated bone marrow (right, representative of two independent experiments). See Fig. 6e, f for similar analysis of basal bone marrow samples. h, Western blots (top) and quantification by densitometry (bottom) showing an increase in S-phase proteins during progression from EEP (P2) to early CEP (P1-CD71low) and late CEP (P1-CD71high). Control 3T3 cells were either cycling or contact-inhibited (non-cycling), as indicated. Western blots are representative of three independent experiments. For gel source images, see Supplementary Fig. 1.
Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size
   Describe how sample size was determined.
   All relevant sample sizes are described in the legend to each figure and/or in the materials and methods section. Sample sizes were determined based on previous experience with each of the experimental systems. Statistical tests were not used to pre-determine sample size.

2. Data exclusions
   Describe any data exclusions.
   Data exclusions are described in detail in the materials and methods section. In the case of scRNA-seq data, contaminating cell types and putative cell doublets were excluded from analysis based on criteria that are described in detail in the method section "Exclusion of contaminating cell types and putative cell doublets". No data was excluded from mouse or cell biology experiments, with the exception of very occasional technical failure in the experimental procedure such as machine failure during a sorting experiment.

3. Replication
   Describe whether the experimental findings were reliably reproduced.
   At least two independent experiments were performed for the results in the manuscript, except for inDrops experiments, which were performed on a single basal bone marrow sample pooled from the bone marrow of 10 mice, or on a single Epo-stimulated bone marrow sample from pooled from the bone marrow of 10 mice; or from a single fetal liver sample pooled from the fetal livers of 10 embryos. Each of the inDrops samples included several thousand cells as indicated in each corresponding figure. All reported data were reproduced reliably.

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.
   Experiments entailed manipulating very similar primary cell samples, derived from inbred mice of similar age, sex and weight, making randomization not applicable.

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   Blinding was not used. Readouts were quantitative and not subject to subjective judgment of investigators.

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

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

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

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

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

- FACS data was analyzed using the FlowJo 9.9.5 software.
- Western blots were analyzed using the Biorad Image Lab Software 6.0
- Data analysis and graphics: GraphPad Prism 7.0a, Excel version 15.32
- Python (version 2.7), using scripts described in methods section and code available at:
  - https://github.com/indrops/indrops
  - https://github.com/AllonKleinLab/SPRING
  - https://github.com/AllonKleinLab/PBA

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

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

There are no restrictions on material availability.
9. Antibodies

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

All antibodies used in this study are commercially available. Where relevant, in addition to validation provided by the supplier, we carried out control experiments with appropriate cell types or biological samples, included in the relevant figures. For example, western blotting for Gata-1 included a positive control of 3T3 cells transduced with the Gata-1 vector. Similarly, western blotting for replication proteins (PCNA, Mcm helicase subunits) included non-dividing (contact inhibited) 3T3 cells as negative controls, and dividing (log-phase expanding) 3T3 cells as positive controls. In the case of flow cytometry antibodies, we relied on supplier validation, and on the expected flow cytometric pattern being obtained for each antibody, since all antibodies in our study are used routinely.

The following antibodies were used in flow cytometry or magnetic bead sorting or analysis:
- anti-CD11b (Clone M1/70 [#557395], BD Biosciences)
- anti-Ly-6G and Ly-6C (Clone RB6-8C5 [#553125], BD Biosciences)
- anti-CD4 (Clone RM4-5 [#553045], BD Biosciences)
- anti-CD8a (Ly-2) (Clone 53-6.7 [#553029], BD Bioscience)
- anti-CD19 (Clone 1D3 [#553784], BD Bioscience)
- anti-TER119 (Clone TER119 [#553672], BD Biosciences)
- CD117- APC Cy7 (Clone 2B8 [#105826], Biolegend)
- TER119- BUV395 (Clone TER-119 [#563827], BD Biosciences)
- CD71- PE Cy7 (Clone R7217 [#113812], Biolegend)
- CD55-AF647 (Clone RIKO-3 [#131806], Biolegend)
- CD105-PE (Clone MJ7/18 [#120408], Biolegend)
- CD150-BV650 (Clone TC15-12F12.2 [#115931], Biolegend)
- CD71- PE Cy7 (Clone R7217, #113812 Biolegend)
- FcεRIα-AF700 (Clone MAR-1,#134323 Biolegend)
- CD41-BV605 (Clone MWReg30 [133921], Biolegend)
- Cd11b-PE Cy5 (Clone M1/70, #101209 Biolegend)
- Ly 6G/C- FITC (Clone RB6-8C5, #53126 BD Biosciences).

The following antibodies were used in western blotting analysis:
- GATA1 (N6, sc-265, Santa Cruz); beta-actin (ab8227, abcam), MCM5 (Bethyl Laboratories, Inc., A300-195A-M), MCM6 (Bethyl Laboratories, Inc., A300-194A), MCM2 (Bethyl Laboratories, Inc., A300-191A), PCNA (PC10) (Santa Cruz, sc-56), IL-17RA/IL-17R (R&D Systems, AF448).

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

b. Describe the method of cell line authentication used.

None. The precise identity of the 3T3-L1 cells is peripheral to the study; they were used as controls in the following way:
1. their ability to become contact inhibited allowed us to compare cycling and non-cycling cells, which was a useful way to obtain controls either expressing or not expressing proteins characteristic of S phase
2. they were used as vehicles, for transduction of GATA-1, in order to determine the location/mobility of the GATA-1 band on western blotting.

3T3-L1 cells were used as controls in some western blots. The entire study was otherwise conducted in primary tissue as indicated (bone marrow or fetal liver)

No

No
Animals and human research participants

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

11. Description of research animals

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

**Mice**

- For scRNA-seq studies: the basal state bone marrow sample (bBM), and for the sorted populations P1 to P5, bone marrow was harvested from 8-week old adult BALB/cJ female mice (Jackson Laboratories, Maine, USA). For the Epo-stimulated adult bone-marrow sample (eBM), 8-week old adult Balb/cJ female mice were injected with Epo (Procrit, Amgen corporation) sub-cutaneously once per 24 hours for a total of 48 hours, at 100 Units/25 g. For the fetal liver sample (FL), BALB/cJ female mice were set up for timed pregnancies, and fetal livers were harvested on embryonic day 13.5.

- For other experiments: BALB/cJ male or female mice, ages 8 to 12 weeks were used.

- For colony formation assays using IL-17RA-deleted mice: To generate the IL-17RA-deleted line, IL-17RA flox/+ mice (El Malki, K. et al. Journal of Investigative Dermatology 133, 441-451, 2013) were bred with CMV-Cre mice (#003465, JAX lab). The generation of il17ra del allele in the F1 generation of il17ra flox/+ × CMV-Cre mating pairs were screened by PCR of tail DNA. To remove the CMV-Cre allele present in the F1 generation, IL-17RA del/+; CMV-Cre−/+ mice were outcrossed with C57BL/6 mice. For colony assays with bone marrow of these mice, control mice were C57BL/6 (Jackson laboratories, stock number 000664).

Policy information about studies involving human research participants

12. Description of human research participants

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

N/A
Flow Cytometry Reporting Summary

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

Data presentation

For all flow cytometry data, confirm that:

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

Methodological details

5. Describe the sample preparation. Please see methods section which details these fully.
6. Identify the instrument used for data collection. LSR II
7. Describe the software used to collect and analyze the flow cytometry data. Acquisition: BD FACSDiva Analysis: Flowjo NB: in Figure preparation, axis labels do not provide fluorochromes, but these are fully detailed for each antibody in the methods section.
8. Describe the abundance of the relevant cell populations within post-sort fractions. Post-sort fractions are indicated on each flow plot as percentages. Sub-populations P1 to P5 were sorted with purities of >90%.
9. Describe the gating strategy used. Gating strategies are fully described in the relevant figures (Fig 2a, 6d)

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