Article

Single-molecule imaging of transcription dynamics in somatic stem cells

https://doi.org/10.1038/s41586-020-2432-4

Received: 27 July 2019

Accepted: 31 March 2020

Published online: 24 June 2020

Check for updates

Justin C. Wheat[1,2], Yehonatan Sella[3], Michael Willcockson[1], Arthur I. Skoultchi[1], Aviv Bergman[3,4,5,6], Robert H. Singer[1,4,7,8,9] & Ulrich Steidl[1,2,10,11]✉

Molecular noise is a natural phenomenon that is inherent to all biological systems[1,2]. How stochastic processes give rise to the robust outcomes that support tissue homeostasis remains unclear. Here we use single-molecule RNA fluorescent in situ hybridization (smFISH) on mouse stem cells derived from haematopoietic tissue to measure the transcription dynamics of three key genes that encode transcription factors: *PU.1* (also known as *Spi1*), *Gata1* and *Gata2*. We find that infrequent, stochastic bursts of transcription result in the co-expression of these antagonistic transcription factors in the majority of haematopoietic stem and progenitor cells. Moreover, by pairing smFISH with time-lapse microscopy and the analysis of pedigrees, we find that although individual stem-cell clones produce descendants that are in transcriptionally related states—akin to a transcriptional priming phenomenon—the underlying transition dynamics between states are best captured by stochastic and reversible models. As such, a stochastic process can produce cellular behaviours that may be incorrectly inferred to have arisen from deterministic dynamics. We propose a model whereby the intrinsic stochasticity of gene expression facilitates, rather than impedes, the concomitant maintenance of transcriptional plasticity and stem cell robustness.

Quantitative, single-cell studies of biological systems have shown that stochasticity is inherent to all cellular processes[1–3]. Owing to low copy-number fluctuations[3–5], spatial and temporal partitioning of reactions[6], and the hard physical bounds that limit efficient feedback control[7], gene expression is inevitably noisy. As such, it is unsurprising that homogeneous transcriptional populations have been challenging—if not impossible—to define[1]. A fundamental question in stem-cell biology is how the robust production of mature cell types arises from these intrinsically stochastic processes.

Haematopoiesis is a paradigmatic stem-cell differentiation model based on the haematopoietic stem cell (HSC) (Fig. 1a). Single-cell RNA sequencing (scRNA-seq) studies have suggested that the gene expression states that underlie terminal branches of the haematopoietic tree arise early and continuously during a multi-step differentiation process through populations of increasingly restricted progenitor cells[8,9]. Transcription factors are thought to have a central role in this process, coordinating the expression of cohorts of target genes during lineage specification. Consequently, determining the magnitude of transcriptional noise in the expression of genes that encode transcription factors is fundamental.

Single-molecule imaging in primary HSPCs

The quantitative evaluation of transcriptional noise requires single-cell techniques with molecular sensitivity. We therefore adapted a single-molecule FISH (smFISH) technique, the gold standard for single-cell RNA analysis, to study transcriptional noise in primary haematopoietic stem and progenitor cells (HSPCs)[10,11]. Owing to the short mRNA length and high (G+C) content of some critical haematopoietic transcription factors, we selected a two-step hybridization strategy to increase the signal-to-noise ratio[12,13] (Fig. 1b, Supplementary Methods 1). We first tested this technique on *PU.1*, which encodes a transcription factor that has essential activating functions in myeloid cell development[14,15], is a repressor of erythroid differentiation[16,17], and the expression of which is deregulated during leukemogenesis[18–20]. Two-step smFISH markedly increased spot intensity, the signal-to-noise ratio, and the number of detectable *PU.1* mRNAs per cell when compared with commercial probe sets (Fig. 1c, Extended Data Fig. 1a, b). We then extended this approach to multiplexed imaging in three channels, enabling the simultaneous detection of three genes in single cells (Extended Data Fig. 1c–e, Supplementary Methods 2).

Because smFISH also enables the direct observation of active transcription sites[21], we first asked how genes that are predicted to be co-regulated with *PU.1* correlated in both mature mRNA counts and in transcriptional activity. We performed multiplexed smFISH for *PU.1* and eight critical haematopoietic genes (transcription factor genes: *Gata1*, *Cebpa*, *Runx1*, *Myb*, *Zfpm1* and *Meis1*; functional genes: *Mpo* and *Gypa*) (Supplementary Table 1) within phenotypically mixed Kit^+Lin^- HSPCs (Extended Data Fig. 1f). Nascent transcription of *PU.1* was higher in

[1]Department of Cell Biology, Albert Einstein College of Medicine, New York, NY, USA. [2]Ruth L. and David S. Gottesman Institute for Stem Cell Research and Regenerative Medicine, Albert Einstein College of Medicine, New York, NY, USA. [3]Department of Systems and Computational Biology, Albert Einstein College of Medicine, New York, NY, USA. [4]Dominick P. Purpura Department of Neuroscience, Albert Einstein College of Medicine, New York, NY, USA. [5]Department of Pathology, Albert Einstein College of Medicine, New York, NY, USA. [6]Santa Fe Institute, Santa Fe, NM, USA. [7]Department of Anatomy and Structural Biology, Albert Einstein College of Medicine, New York, NY, USA. [8]Gruss-Lipper Biophotonics Center, Albert Einstein College of Medicine, New York, NY, USA. [9]Janelia Research Campus of the HHMI, Ashburn, VA, USA. [10]Department of Medicine (Oncology), Albert Einstein College of Medicine-Montefiore Medical Center, New York, NY, USA. [11]Albert Einstein Cancer Center, Albert Einstein College of Medicine, New York, NY, USA. ✉e-mail: ulrich.steidl@einsteinmed.org

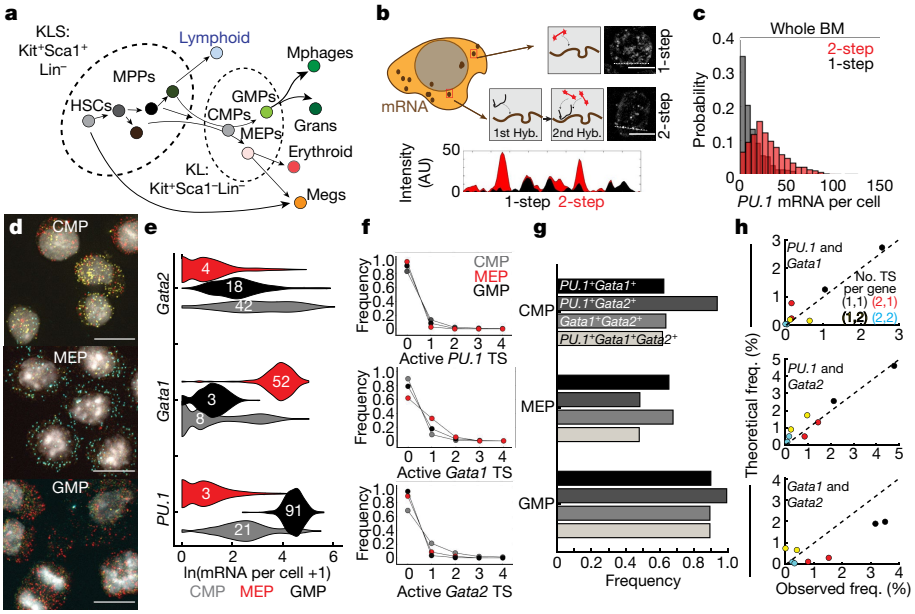


Fig. 1 | Stochastic bursting of mRNAs drives co-expression of antagonistic transcription factors in HSPCs. a, Schematic of haematopoietic hierarchy. HSCs, haematopoietic stem cells; grans, granulocytes; mphages, macrophages; MPP, multi-potent progenitors. **b**, Description of smFISH using a two-step hybridization (hyb.) method. The bottom graph shows line plots of the signal above the background. **c**, Quantification of *PU.1* molecules per bone marrow (BM) mononuclear cell using 1-step or 2-step smFISH. **d**, Filtered images of CMP, GMP and MEP cells stained by smFISH for *PU.1* (Cy5, red pseudocolour), *Gata1* (Alexa Fluor 594, cyan pseudocolour), and *Gata2* (Cy3, yellow pseudocolour). Scale bars, 10 µm. DNA is shown in grey pseudocolour. **e**, Violin plots (area-normalized) of the natural log normalized (mRNA per cell + 1) distribution for each gene. The numbers overlaid are the mean copy number per cell (CMP, $n = 3{,}174$; GMP, $n = 364$; MEP, $n = 1{,}113$). **f**, Burst frequency of each gene in each HSPC subpopulation. TS, transcription site. **g**, Frequency of cells co-expressing *PU.1*, *Gata1* and *Gata2*. **h**, Comparison of observed co-bursting frequencies versus theoretical frequencies derived from statistical independence. The colour indicates which combination of bursting patterns is being tested—for example, (1,2) in the top panel means the frequency of cells with 1 active *PU.1* site and 2 active *Gata1* sites. The dashed line is $y = x$. Data in **d**–**h** are derived from 2 independent experiments for CMPs and MEPs and 1 experiment for GMPs.

cells in which the expression levels of *PU.1* were high (a *PU.1*^{high} state), as expected (Extended Data Fig. 1g, h). Furthermore, cells in the *PU.1*^{high} state showed increased nascent transcription of the myeloid lineage genes *Cebpa*, *Mpo* and *Myb*, and reduced nascent transcription of the erythroid genes *Gata1*, *Zfpm1* and *Gypa* (Extended Data Fig. 1i, j), as well as *Meis1*; this was also consistent with expectation. We found minimal change in the expression of *Runx1*. These experiments demonstrate the utility of smFISH in studying transcription in primary HSPCs.

Next we compared the sensitivity of smFISH with that of scRNA-seq, by comparing mRNA detection of the seven aforementioned transcription factor genes by smFISH with five scRNA-seq datasets of comparable HSPCs[9,22–25]. For all genes tested, there was a marked increase in the number of non-expressing cells in the scRNA-seq datasets compared with the smFISH results (Extended Data Fig. 2a). We then calculated the Gini coefficient for each gene, which quantifies population dispersion of a variable of interest[11]. Because capturing the underlying population dispersion for an mRNA is essential for assigning transcriptional states, underestimating this metric implicitly limits information about gene regulation. Gini coefficients range from a value of 0 (equal distribution of gene-expression values) to a value of 1 (a minority of cells with signal greater than 0). For six out of seven genes tested, the Gini coefficient was lower when determined by smFISH than by scRNA-seq (Extended Data Fig. 2b). The sole exception was *Zfpm1*, for which one out of five scRNA-seq studies showed a similar calculated Gini index to that obtained from smFISH. We then calculated Gini coefficients for a larger set of other transcriptional regulatory genes from scRNA-seq data (Supplementary Table 2). Consistent with the findings of our initial test set of transcription factor genes, the majority of genes in this list had Gini indices greater than 0.8 (Extended Data Fig. 2c). Furthermore, we found that these sensitivity restrictions substantially affect both unsupervised clustering of transcriptional states and post hoc analyses of the pairwise dependencies in the expression of transcription factors (Extended Data Fig. 2d–f, Supplementary Fig. 2a, Supplementary Table 3, Supplementary Discussion 1). As such, scRNA-seq may be fundamentally incapable of providing quantitative estimates of transcriptional noise during haematopoiesis.

Co-expression of *PU.1* and *Gata* genes in HSPCs

We then used smFISH to evaluate the role of stochasticity in a central transcriptional network between *PU.1* and the *Gata* transcription factor genes *Gata1* and *Gata2*. PU.1 and GATA1 are critical to differentiation along the granulocyte–monocyte and erythrocyte lineages, respectively, and the direct interaction between these transcription factors through an antagonistic toggle switch was the original model for fate decisions along the granulocyte–monocyte or erythrocyte lineages[16,17]. *Gata2* is abundant in early HSPCs and may function similarly to *Gata1* in these cells by antagonizing the function of *PU.1*, albeit at lower potency[26]. Additionally, *Gata2* primes HSPCs to upregulate *Gata1* during terminal erythropoiesis, after which *Gata1* is thought to shut off *Gata2* in a phenomenon described as the 'Gata switch'[27,28]. Nevertheless, recent scRNA-seq studies have either failed to detect progenitors that co-express *PU.1* and *Gata1*[23], or detected co-expression in only a small minority of cells[8,9,22], which has called into question the validity of such a model in directing myeloid–erythroid fate decisions.

We isolated three immunophenotypically defined populations—granulocyte/monocyte progenitors (GMPs), megakaryocyte/erythrocyte progenitors (MEPs) and common myeloid progenitors (CMPs)—and assessed the expression of these transcription

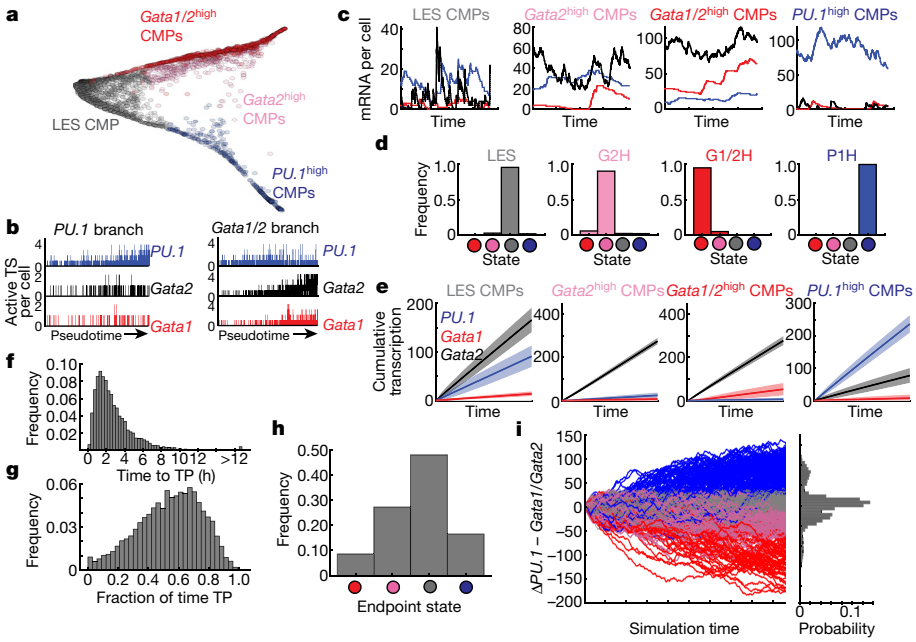


Fig. 2 | Inferred dynamics of the *PU.1/Gata1/Gata2* network in CMPs. **a**, Diffusion pseudotime mapping of CMPs, coloured according to transcriptional state. **b**, Transcription site bursting frequency with increasing pseudotime along each branch. **c**, Single trajectories of three-gene stochastic simulation. **d**, Stability of transcriptional states using inferred parameters. **e**, Average cumulative mRNA produced during the simulation. The line indicates the mean among simulations; shaded regions are ±s.d. $n = 10{,}000$. **f**, **g**, Time-dependent behaviour of simulated cells in the LES parameter regime, initialized at 0 mRNAs for all three genes at $t = 0$. **f**, Histogram covering the time from the start of the simulation to the first time point of instantaneous co-expression—that is, triple positive (TP). All first TP events >12 h were pooled together. **g**, Histogram of total simulation time spent in the triple-positive state (mean, 56.8%; s.d., 20.6%, $n = 10{,}000$). **h**, **i**, Analysis of noise-derived transitions between states and efficacy of system evolution from LES. **h**, Frequency of each endpoint state after 12 h of simulation time, initialized in the LES state. $n = 10{,}000$. **i**, Behaviour of simulation trajectories over time, coloured on the basis of endpoint state. Right, the marginal distribution of endpoints.

factors by smFISH (Fig. 1d, Supplementary Fig. 1). *Gata1* was high in MEPs and low in GMPs, while *PU.1* was high in GMPs and low in MEPs; *Gata2* was highest in CMPs (Fig. 1d, e, Extended Data Fig. 3). Notably, in all instances—apart from *PU.1* in GMPs and *Gata1* in MEPs—we found that mRNA count distributions were positively skewed, with the majority of the probability mass below 50 copies of mRNA per cell. Frequency distributions of this type are typical of mRNAs produced in infrequent bursts of transcription[29]. Consistently, all genes were infrequently 'ON', even in high-expressing cell types (Fig. 1f).

Given the infrequency of active sites and the relatively low copy number of each gene, we next assessed the frequency of co-expression of these genes. The majority of CMPs expressed *PU.1* (97%), *Gata2* (96%) and *Gata1* (64%) (Extended Data Fig. 3c). Notably, greater than 60% of CMPs had at least one mRNA for all three genes, as well as 45% of MEPs and 89% of GMPs (Fig. 1g; see Extended Data Fig. 4 and Supplementary Discussion 2 for discussion of false positives in smFISH).

We next asked whether CMPs were still actively transcribing *PU.1* and either of the *Gata* genes, or whether co-transcription of these factors was precluded at this stage of differentiation. To test this, we used the fact that if nascent transcription of *PU.1* and *Gata1* or *Gata2* were mutually exclusive, the empirical frequency of CMPs with simultaneous transcription sites, $f_{PU.1^+Gata1^+}$, should be lower than the frequency predicted by statistically independent firing, $f_{PU.1^+} \times f_{Gata1^+}$. On the contrary, we found that both $f_{PU.1^+Gata1^+}$ and $f_{PU.1^+Gata2^+}$ were essentially indistinguishable from those predicted by statistically independent bursting (Fig. 1h). Additionally, $f_{Gata2^+Gata1^+}$ was approximately 1.5- to twofold higher than $f_{Gata2^+} \times f_{Gata1^+}$, consistent with the proposed model of *Gata* gene co-expression during erythropoiesis. These findings indicate that mutually exclusive transcription of the antagonistic transcription factors *PU.1* and *Gata1* or *Gata2* does not occur in CMPs.

Stochastic transitions to transcriptional termini

We then performed stochastic simulations using the transcriptional parameters inferred from our CMP data to model the transcriptional behaviour of each gene over time[21,30]. To refine our parameter fitting, we assigned CMPs to four transcriptional states: a *PU.1*^{high}*Gata1/2*^{low} state (P1H); a *Gata1/2*^{high}*PU.1*^{low} state (G1/2H); a *Gata2*^{high} state (G2H); and a state with low expression of all three genes (LES CMPs) (Extended Data Fig. 5a, b, Methods). Ordering these states with diffusion pseudotime estimation[31] (Fig. 2a) identified two branches emanating from the LES CMP cluster. Consistent with our previous analysis, although each branch in the pseudotime plot had differential transcriptional activity, active transcription sites for the 'opposing' transcription factor were still detected even late in pseudotime along a given branch (Fig. 2b, Extended Data Fig. 5c). We then inferred the transcriptional rate parameters for each state using the smFISH data (Supplementary Table 4, Supplementary Methods 4). Single-cell trajectories simulated using parameters for a given state closely approximated the transcriptional behaviour of each state (Fig. 2c), and were extremely stable (Fig. 2d). We then used these simulations to infer the cumulant number of nascent mRNAs produced in each state over a time frame typical of the lifespan of CMPs in vitro[32]. The majority of trajectories transcribed hundreds of copies of *Gata2* over this time period, irrespective of the transcriptional state (Fig. 2e). Simulated LES CMPs and G2H CMPs also transcribed between 20 and 100 mRNAs for *Gata1* and *PU.1*, respectively. On average, LES cells were predicted to contain mRNAs of all three genes after just two hours of simulation time, and greater than 99% of cells were 'triple positive' at some point during the 12-h simulation window (Fig. 2f). Furthermore, the majority of trajectories were triple positive for over half of the simulation timeframe (Fig. 2g).

We then asked if bifurcation into P1H and G1/2H states could occur stochastically from the LES state. Indeed, although the parameter set

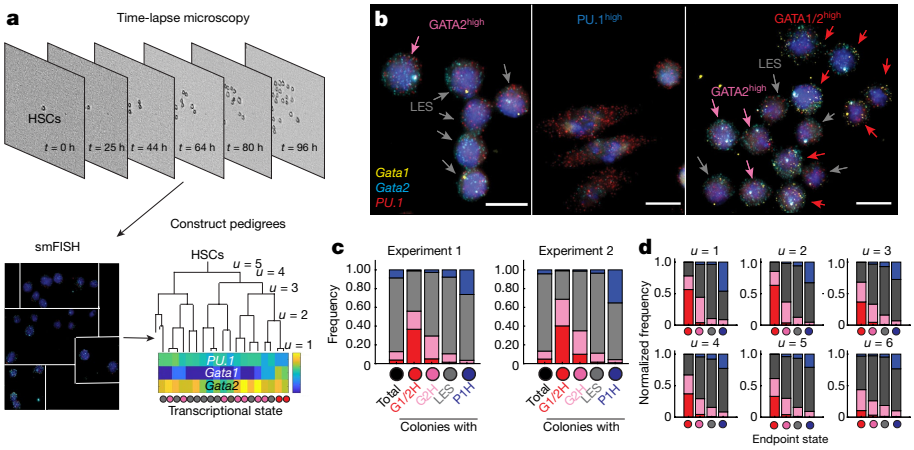


Fig. 3 | Transcription-state correlation among clonal progeny of single HSCs. a, Schematic of experimental workflow. The smFISH image is a stitched composite of four separate fields of view. The heat map associated with the pedigree represents the ln(mRNA per cell + 1). Coloured spheres indicate the assigned transcriptional state of the cell. **b**, Representative images of cells in each endpoint state under study (number of experiments = 2). Scale bars, 10 μm. **c**, Frequency of states within mixed colonies conditional on the presence of each state. 'Total' represents the frequency of states in all cells analysed at the endpoint. The empirical distribution of the four HSPC states at the 96-h endpoint was 2.9% (G1/2H), 14.50% (G2H), 6.9% (P1H) and 74.8% (LES). Experiment 1, 34 colonies; experiment 2, 83 colonies. **d**, Frequency of state pairs at generational distances $u = 1$ to $u = 6$ as indicated in **a**, normalized to the frequency of each state. Endpoint states are demarcated by coloured circles under each bar plot.

of each state generated stable trajectories that largely maintained their initial state assignment, rare transitions to other states did occur (Fig. 2d). Therefore, we repeated simulations by first initializing cells in the LES state, and through fluctuations alone allowed cells to transition to other states in which they would then adopt new transcriptional parameters. Of trajectories initialized in the LES state, 9% and 18% ended up in either the G1/2H or the P1H terminus after one CMP lifetime, respectively, whereas 25% of trajectories ended up in the G2H state (Fig. 2h). Trajectories ending in the terminal G1/2H and P1H states frequently fluctuated in and out of the LES and G2H states (Fig. 2i). Moreover, changes in nascent transcription rates were required for cells to reach both termini (Extended Data Fig. 5d–f). These analyses indicate that, although transcriptional noise drives co-expression of antagonistic transcription factors, stochastic and reversible transitions of noisy states can still efficiently bifurcate into *PU.1*^{high} and *Gata*^{high} expressing states.

Mapping HSC state correlations by pedigree analysis

Although the above results suggest considerable transcriptional stochasticity in CMPs, whether such phenomena occur in HSCs is a critical question. Moreover, the effect such processes have on the transcriptional state dynamics of the *PU.1*–*Gata1* network in HSCs is currently debated[32–34], and 'transcriptional priming' has been suggested as putatively limiting the transcriptional states an HSC and its descendants can occupy[9,35,36].

We first asked whether HSCs co-expressed *PU.1* and the *Gata* genes. HSC and their early progeny showed robust expression of all three genes at a similar level to that of CMPs, with greater than 99% of cells expressing *PU.1* and *Gata2* and 55% co-expressing all three mRNAs (Extended Data Fig. 6a–c).

Next, to understand how the temporal dynamics of these genes are coordinated, we used kin correlation analysis (KCA)—an experimental approach that uses the information embedded in pedigrees to infer the dynamics of transcriptional state transitions[37]. To that end, we followed HSCs for 96 h ex vivo, constructed pedigrees from each HSC, and used smFISH to assign transcriptional states to cells (Fig. 3a, Extended Data Fig. 7, Supplementary Fig. 2b; see Methods and Supplementary Methods 5 for details of state assignments). In addition to the four subpopulations identified in CMPs (Fig. 3b) (LES, G1/2H, G2H and P1H), we also detected some cells in a megakaryocytic state (Megs) that had hundreds of copies of each of the three mRNAs and were polyploid (Extended Data Fig. 7a), as well as rare (0.74%) cells with macrophage-like morphology and very high *PU.1* levels. We excluded these cell populations to focus on more immature HSPCs.

First, we determined whether individual HSCs could generate progeny in multiple states. Twenty-seven out of 117 colonies contained only one predominant state type: 5 out of 117 were G1/2H-dominant, 2 out of 117 were P1H-dominant, and 21 out of 117 were LES-restricted (Extended Data Fig. 8a). The frequency of colonies with any combination of 2, 3, or 4 or more states was around 45%, around 25% and around 3%, respectively. Of mixed colonies, 25% had at least one G1/2H cell and 41% had at least one P1H cell. All other colonies were composed of mixtures of G2H and LES. To determine which combination of states could derive from a single clone, we calculated the frequency of states within mixed colonies that were conditional on the presence of a cell in each state (Fig. 3c). Although no two states were mutually exclusive in this analysis, the frequency of finding a colony with both G1/2H and P1H states was low (3 out of 117). HSC colonies that produced any G1/2H progeny showed a tenfold reduction in the frequency of P1H cells, a tenfold increase in the frequency of G1/2H cells, and a 1.5-fold increase in G2H cells. Similarly, colonies with any endpoint progeny in the G2H state had reduced frequencies of cells in the P1H and LES cell states, while the frequency of cells in the G1/2H state was increased nearly twofold. Conversely, clones producing P1H cells had a threefold reduction in G1/2H cells and a fourfold reduction in G2H cells.

Stochastic and reversible HSC transcription dynamics

One scenario that could account for such behaviour is an irreversible switch in the transcriptional kinetics that arises early in the pedigree. In such a case, cells at close generational distances (for example, sister cells) would be expected to be in the same transcriptional state. However, we found that P1H and G1/2H states were paired with LES and G2H states even at recent generational divides, including sister cells (lineage distance $u = 1$) (Fig. 3d). These results indicate that transitions to a high-expressing state either occurred irreversibly but late, or were infrequent and reversible.

To discriminate between these mutually exclusive hypotheses, we used KCA, which uses the correlation between endpoint transcription

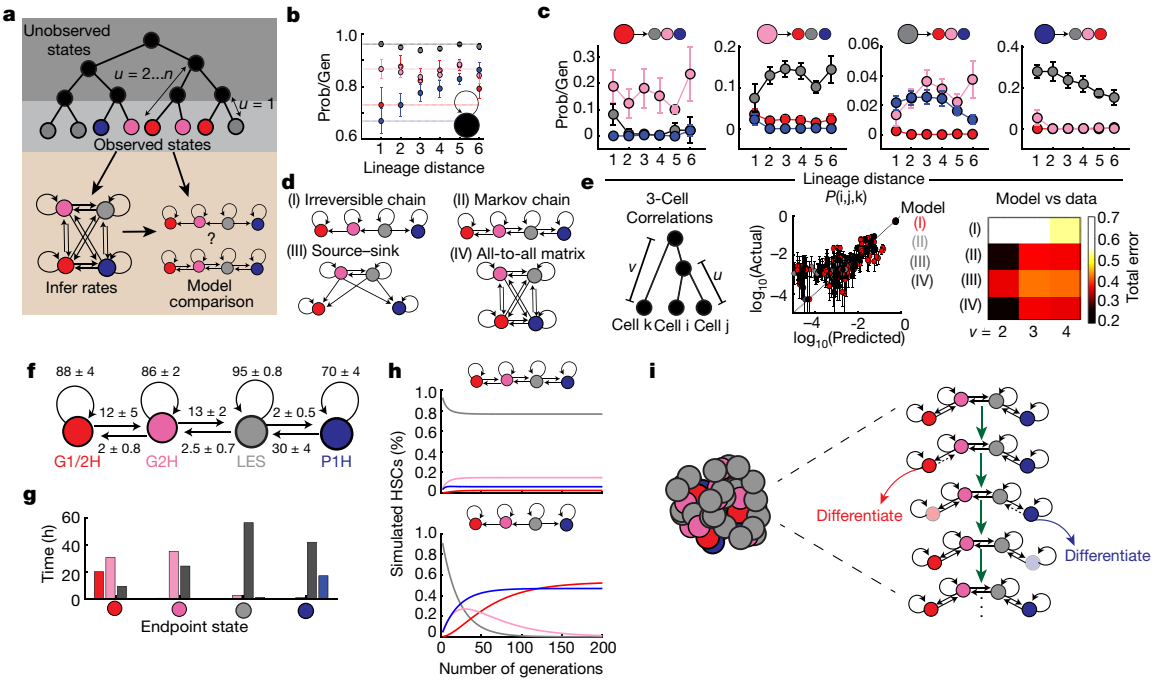


Fig. 4 | Stochastic and reversible HSC transcription state dynamics. **a**, Schematic of KCA. **b**, **c**, Inferred state persistence (**b**) and state transition (**c**) rates, given as probability per generation for each lineage distance. Circles with error bars are mean inferred rate with standard error derived by bootstrapping through data ($n = 5{,}000$). Dotted horizontal lines in **b** are the rates at $u = 1$. **d**, **e**, Using three-point state frequencies to compare models. **d**, Schematic of tested state transition models. **e**, Left, schematic of three-point state frequencies. Middle, observed versus theoretical three-point frequencies as predicted by each model. Each circle with error bars is the mean experimental three-state frequency (y axis) and inferred average three-state frequency (x axis) at a given distance. The error bars are the experimental standard error derived by bootstrapping ($n = 1{,}000$). Right, total error between theory and observed frequencies at $v = 2{:}4$ for each model. Models with irreversible edges between states have higher error—that is, less predictive value—than those with reversible edges. **f**, Transition probabilities (mean ± s.e.m.) per generation for the inferred Markov chain. **g**, Average fraction of time spent in each state for a given endpoint state, conditional on the structure of the pedigree and state distribution of progeny. **h**, State frequencies over generational time when reversible (top) and irreversible (bottom) dynamics connect transcription states, initialized in the LES state. Curve colours correspond to each state as in **c**. **i**, Proposed model of reversible transcription state transitions connecting *PU.1* and *Gata* states in early HSPCs.

states and lineage distance between cells in a pedigree to infer the transition rates of those states (Fig. 4a). Using the pedigree and smFISH data, we first determined a transition matrix across all generational distances and across all edges (Fig. 4b, c). The inferred transition rates between any two states were relatively low compared with the probability of retaining the state of the parent cell (Fig. 4b, c), consistent with the observation that most cells had not transitioned to the P1H or G1/2H state by the endpoint of the experiment. Additionally, we noted that some transitions had little to no probability per generation—for example, direct transitions from P1H to G1/2H or vice versa had approximately 0% probability. Moreover, entering the G2H state seemed to be a prerequisite for entering G1/2H. The inferred transition probabilities were robust to a range of mRNA cutoffs between different states, suggesting that these transitions are not artefacts of noise across an arbitrary cutoff (Extended Data Fig. 9). Moreover, we found no evidence of partitioning asymmetries of mRNAs during division, excluding the possibility that such phenomena influenced the inferred transition probabilities (Extended Data Fig. 10).

We next used these transition probabilities to model a range of state transition behaviours, from a fully irreversible chain of commitment (Fig. 4d, model I) to a fully connected network (Fig. 4d, model IV). We compared the predictive power of these models by determining the error between the three-cell state frequencies predicted by each model and those observed in the experiment (Fig. 4e, Methods). At all generational distances tested, for both the Markov chain (model II) and the fully connected model (model IV) there were approximately 100% and 30% reductions in the predicted three-cell state frequency error when compared with the irreversible (model I) and partially irreversible (model III) models, respectively (Fig. 4e). Moreover, the Markov chain performed better at lower generational distances v (for example, $v = 2$). Overall, among the models tested, state transition models that contained reversible transitions to the P1H and G1/2H states outperformed those with irreversible transitions, and the Markov chain model best captured the underlying state transitions (Fig. 4f).

Finally, we aimed to determine the state histories of a cell given its current transcriptional state and the state of its clonal relatives. We found that the majority of time along any trajectory was spent in the LES and G2H states, including those generating a P1H or G1/2H endpoint state (Fig. 4g). As such, a Markov chain governed by these parameters can lead to priming-like behaviours in clonal descendants of single HSCs without necessitating early, irreversible transitions of states or noiseless regulation of transcription. Of note, this analysis indicates that the current transcriptional state of a cell—as defined by the three genes *PU.1*, *Gata1* and *Gata2*—may not be fully predictive of the past or future states visited by the ancestors or descendants of that cell, respectively, even though it may bias the probability distribution of obtainable states in the short term.

Discussion

How robust cellular phenotypes arise from intrinsically noisy processes is a question of central importance to the study of tissue morphogenesis and organismal homeostasis. Although 'playing dice' with gene expression networks may seem counterproductive, such strategies could be evolutionarily advantageous for tissue homeostasis (Fig. 4h, Supplementary Discussion 3). Indeed, such systems have the advantage

of maintaining a temporally stable probability of cells in every available transcriptional state (Supplementary Methods 6) without requiring complex regulatory measures to facilitate such behaviour, all of which will be similarly subject to molecular noise[7] (Supplementary Discussion 3).

Here we have attempted to quantitatively address the question of noise in the expression of transcription factor genes in primary HSPCs, using single-molecule imaging and the quantitative analysis of pedigrees. Our results indicate that antagonistic transcription factors are co-expressed in the majority of HSPCs, and that stochastic transitions between the transcriptional states defined by these genes are the probable basis of the dynamics of the system (Fig. 4i). As these dynamics generate stability by leveraging the physically intrinsic noise of gene expression, the conclusions from these studies may reflect a central and unifying principle underlying the properties of stem and progenitor cells that are central to the evolution of metazoan life. As such, determining whether these reported phenomena predominate in other tissue systems will be critical to developing a quantitative understanding of organismal homeostasis and, consequently, the pathobiology of diseases originating in or influenced by tissue-resident stem-cell compartments.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2432-4.

1. Levsky, J. M. & Singer, R. H. Gene expression and the myth of the average cell. *Trends Cell Biol.* **13**, 4–6 (2003).
2. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
3. Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science* **304**, 1811–1814 (2004).
4. Bar-Even, A. et al. Noise in protein expression scales with natural protein abundance. *Nat. Genet.* **38**, 636–643 (2006).
5. Gandhi, S. J., Zenklusen, D., Lionnet, T. & Singer, R. H. Transcription of functionally related constitutive genes is not coordinated. *Nat. Struct. Mol. Biol.* **18**, 27–34 (2011).
6. Huh, D. & Paulsson, J. Random partitioning of molecules at cell division. *Proc. Natl Acad. Sci. USA* **108**, 15004–15009 (2011).
7. Lestas, I., Vinnicombe, G. & Paulsson, J. Fundamental limits on the suppression of molecular fluctuations. *Nature* **467**, 174–178 (2010).
8. Olsson, A. et al. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537**, 698–702 (2016).
9. Tusi, B. K. et al. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018).
10. Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H. Visualization of single RNA transcripts in situ. *Science* **280**, 585–590 (1998).
11. Torre, E. et al. Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA FISH. *Cell Syst.* **6**, 171–179.e5 (2018).
12. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
13. Tsanov, N. et al. smiFISH and FISH-quant – a flexible single RNA detection approach with super-resolution capability. *Nucleic Acids Res.* **44**, e165 (2016).
14. Chen, H. M., Pahl, H. L., Scheibe, R. J., Zhang, D. E. & Tenen, D. G. The Sp1 transcription factor binds the CD11b promoter specifically in myeloid cells in vivo and is essential for myeloid-specific promoter activity. *J. Biol. Chem.* **268**, 8230–8239 (1993).
15. Koschmieder, S., Rosenbauer, F., Steidl, U., Owens, B. M. & Tenen, D. G. Role of transcription factors C/EBPα and PU.1 in normal hematopoiesis and leukemia. *Int. J. Hematol.* **81**, 368–377 (2005).
16. Rekhtman, N., Radparvar, F., Evans, T. & Skoultchi, A. I. Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells. *Genes Dev.* **13**, 1398–1411 (1999).
17. Zhang, P. et al. PU.1 inhibits GATA-1 function and erythroid differentiation by blocking GATA-1 DNA binding. *Blood* **96**, 2641–2648 (2000).
18. Rosenbauer, F. et al. Acute myeloid leukemia induced by graded reduction of a lineage-specific transcription factor, PU.1. *Nat. Genet.* **36**, 624–630 (2004).
19. Steidl, U. et al. Essential role of Jun family transcription factors in PU.1 knockdown-induced leukemic stem cells. *Nat. Genet.* **38**, 1269–1277 (2006).
20. Will, B. et al. Minimal PU.1 reduction induces a preleukemic state and promotes development of acute myeloid leukemia. *Nat. Med.* **21**, 1172–1181 (2015).
21. Skinner, S. O. et al. Single-cell analysis of transcription kinetics across the cell cycle. *eLife* **5**, e12175 (2016).
22. Giladi, A. et al. Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat. Cell Biol.* **20**, 836–846 (2018).
23. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
24. Nestorowa, S. et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016).
25. Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
26. Chou, S. T. et al. Graded repression of *PU.1/Sfpi1* gene transcription by GATA factors regulates hematopoietic cell fate. *Blood* **114**, 983–994 (2009).
27. Doré, L. C., Chlon, T. M., Brown, C. D., White, K. P. & Crispino, J. D. Chromatin occupancy analysis reveals genome-wide GATA factor switching during hematopoiesis. *Blood* **119**, 3724–3733 (2012).
28. Grass, J. A. et al. GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proc. Natl Acad. Sci. USA* **100**, 8811–8816 (2003).
29. Singer, Z. S. et al. Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Mol. Cell* **55**, 319–331 (2014).
30. Gillespie, D. T. A general method of numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434 (1976).
31. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
32. Hoppe, P. S. et al. Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature* **535**, 299–302 (2016).
33. Buggenthin, F. et al. Prospective identification of hematopoietic lineage choice by deep learning. *Nat. Methods* **14**, 403–406 (2017).
34. Strasser, M. K. et al. Lineage marker synchrony in hematopoietic genealogies refutes the PU.1/GATA1 toggle switch paradigm. *Nat. Commun.* **9**, 2697 (2018).
35. Arinobu, Y. et al. Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages. *Cell Stem Cell* **1**, 416–427 (2007).
36. Laslo, P. et al. Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell* **126**, 755–766 (2006).
37. Hormoz, S. et al. Inferring cell-state transition dynamics from lineage trees and endpoint single-cell measurements. *Cell Syst.* **3**, 419–433.e8 (2016).

Methods

All reagents used in these studies are listed with catalogue number in Supplementary Table 5 and 6.

Animal husbandry

Male and female C57/BL6 mice (6–10 weeks old) were purchased from Jackson Laboratories and housed in animal facilities at the Albert Einstein College of Medicine. All experiments were approved by the Institutional Animal Care and Use Committee of the Albert Einstein College of Medicine (2016-1003). All procedures were performed in accordance with guidelines from the Institutional Animal Care and Use Committee of the Albert Einstein College of Medicine. The number of animals used was not specified at the beginning of the study, and randomization and blinding were not performed.

Cell lines

HPC-7 cells were passaged in IMDM +5% fetal bovine serum, 1% penicillin/streptomycin, 1% sodium bicarbonate, 74.8 μM monothioglycerol and recombinant mouse (rm) SCF (50 ng/ml). The HPC-7 cell line was originally provided by O. Abdel-Wahab. Cells were not authenticated or tested for mycoplasma.

Primary HSPC cultures

Primary HSPC were isolated by cell sorting on a Moflo Astrios EQ (Beckman Coulter). Kit^+Lin^- (KL) populations (CMPs, MEPs and GMPs) were grown on retronectin coated (40 μg/ml) #1.0 glass, 35mm^2 MatTek dishes in IMDM with 1% penicillin/streptomycin, 10% FBS and supplemented with recombinant mouse (rm) SCF (100 ng/ml), rmTPO (100 ng/ml), rmIL-3 (10 ng/ml), rmIL-6 (10 ng/ml), and recombinant human (rh) EPO (2 IU/ml) and GM-CSF (10 ng/ml). M-CSF (10 ng/ml) and G-CSF (10 ng/ml) was supplemented to GMP cultures. Bulk KL cells used in Extended Data Figs. 1 and 2 were grown in suspension in a single well of a 24-well plate.

Cells were grown for approximately 12–16 h ex vivo to allow for full recovery from sorting before analysis with smFISH. HSCs in Extended Data Fig. 6 were grown for 72 h on retronectin coated MatTek dishes, as above, in StemSpan SFEM media with 1% penicillin/streptomycin and recombinant mouse (rm) SCF (100 ng/ml), mTPO (100 ng/ml), rmIL-3 (10 ng/ml), rmIL-6 (10 ng/ml), and rhEPO (2 IU/ml). Cells were maintained at 37 °C and 5% CO_2.

Onstage culture

For time-lapse imaging, sorted HSCs were seeded on 35 mm^2 MatTek dishes coated with 10 μg/ml anti-CD43 biotin instead of retronectin in order to reduce cell movement and cell loss and/or misidentification during the experiment[38]. Cultures were maintained at 37 °C with humidity and 5% CO_2/95% premixed air using the Evos FL2 Auto Onstage Incubator.

Flow cytometry and cell sorting

Mice (5–10 per experiment) were euthanized by CO_2 asphyxiation followed by cervical dislocation. Sternum, tibiae, femurs, pelvic bones and vertebrae were isolated, pooled, and crushed with a mortar and pestle on ice in MACS buffer (PBS, 1% FBS, 1 mM EDTA) and filtered through a 70-μm filter. Red blood cells and granulocytes were then removed through density centrifugation over a 5 ml Histo-Paque Ficoll Gradient. After extensive washing of the buffy coat, cells were then lineage-depleted using 1:1,000 dilution of anti-mouse B220, CD19, CD4, CD8, Gr-1, CD11b, Ter119 and CD127, all biotinylated, on ice for 25 min. Cells were washed and then stained with triple-washed anti-IgG magnetic beads (Untouched Mouse T Cells Kit, Thermo Fisher) on ice for 30 min. Cells were washed and then depleted of lineage-positive cells by passing through a magnetic separation column (MACS LD Column, Militenyi) loaded on a QuadraMACS magnet (Militenyi). Lineage-negative cells were then stained for 30 min on ice with anti-CD150, anti-CD34, anti-KIT, anti-Sca1 and anti-CD48 (all 1:250) and anti-CD16/32 (1:500) with Streptavidin Pacific Orange (1:1,000). Cell populations were sorted on 4-way purity mode into IMDM, 5% FBS, 1% penicillin/streptomycin. See Supplementary Methods for gating strategy.

Poly-L-lysine coating of #1.0 12 mm coverslips

To prepare poly-L-lysine-coated coverslips for immobilization of suspension cells (HPC-7, Kit^+Lin^- progenitors, and whole bone marrow), 12 mm #1 Coverslips were first boiled in 0.5 M HCl for 30 min, washed extensively in double distilled water and stored in 70% ethanol. Coverslips were then coated for 5 min with 0.01% poly-L-lysine, followed by two washes with water and air-dried for 20 min. Coverslips were then transferred to a 24-well dish on ice for cell immobilization and subsequent smFISH staining. Cell aliquots (20 μl) of around 10,000 cells per 100 μl were dotted and spread onto the coverslip and the cells were allowed to settle on ice for 20 min. Unstuck cells were then washed away with two PBS washes before fixation and smFISH staining.

Probe design

To design mRNA-specific targeting probes for sequential smFISH, mRNA sequences—including 5′ and 3′UTRs for each gene—were imported into Oligo7 software. 30mer targeting sequences were identified as follows, with a minimum of 10bp between successive probes: GC content 50–60%; ΔG (free energy) of duplexes greater than −0.1 kJ/mol; ΔG of hairpin formation greater than −0.1 kJ/mol.

Putative sequences were then screened for off-target activity using Blast (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Selected sequences were then concatenated on the 5′ and 3′ end with flanking readout 20mer sequences, generating a final 'primary probe' as a 70mer. Probes were then ordered in 100 nmol quantities from Thermo Fisher or IDT. Individual probes were resuspended at 100 μM concentration, mixed in equal proportions to 10 μM final concentration of each probe, and stored at −20 °C. Stock solutions were diluted in Ultrapure water to 200 ng/μl for working stocks.

For the design of *Mpo*, *Myb* and direct *PU.1* 18–20mer probes, mRNA sequences were imported into the Stellaris Probe Designer tool (LGC Biosearch) with masking level 5, oligo length 20 and minimum spacing of 2nt. Commercial probes were used at 10 nM final concentration.

Sequential smFISH for *PU.1*, *Gata1*, *Gata2*, *Cebpa*, *Runx1*, *Meis1*, *Zfpm1* and *Gypa*

Cells were fixed in 3.2% PFA (Electron Microscopy Sciences) and diluted in PBS with 1 mM $MgCl_2$ (PBSM) at room temperature for ten minutes. Cells were then washed with 2 ml cold PBSM with 10 mM glycine. Cells were then permeabilized on ice for 20 min in PBSM with 0.1% Triton X-100 and 2 mM vanadyl ribonucleoside complex (VRC). After washing with PBSM, cells were then incubated at room temperature with prehybridization-30 buffer (prehyb-30; 30% formamide, 2X SSC). Cells were then stained overnight at 37 °C with hybridization buffer consisting of 10% dextran sulfate, 30% formamide, 2X SSC, 2 mM VRC, 10 μg/ml sheared ssDNA from salmon sperm, 10 μg/ml *E. coli* tRNA, 10 μg/ml molecular grade bovine serum albumin, and 200 ng each of 70mer primary probe mixes. Cells were then washed twice for 20 min at 37 °C with prehyb-30, and once with 2X SSC. Cells were then post fixed in 1% PFA in PBSM for 5 min, followed by two washes in 2X SSC. Primary stained cells were then washed with prehyb-10 (10% formamide, 2X SSC) for 10 min at 37 °C and stained with 10% dextran sulfate, 10% formamide, 2X SSC, 2 mM VRC, 10 μg/ml sheared ssDNA from salmon sperm, 10 μg/ml *E. coli* tRNA, 10 μg/ml molecular grade bovine serum albumin, and 10 ng each of 20mer readout probes for each gene for 3 h at 37 °C. Reactions were then washed twice for 10 min in prehyb-10, followed by a final wash in 2X SSC. Cells were then mounted in Prolong Diamond Antifade reagent plus DAPI. For cells grown on MatTek dishes, the mounting was performed by laying a 12 mm #1.0 coverslip onto the central glass well of the dish; for cells immobilized

on a coverslip, the coverslip was first blotted on filter paper to remove excess moisture and then inverted onto a drop of Antifade on a glass slide. See Supplementary Table 1 for probe sequences.

Probe labelling with Cy3, Cy5 and Alexa Fluor 594

Secondary 'readout' probes were purchased from Thermo Fisher with 5′ C5 amine and 3′ C7 amine modifications. Five micrograms of each readout probe was then coupled to the appropriate fluorescent dye according to the manufacturer's specifications. After labelling, probes were extracted from excess dye by use of a Qiagen Nucleotide Removal Kit, resuspended in Ultrapure water, and stored at −20 °C. Labelling efficiency was determined using Beer's law. Only fluorescently labelled probes with more than 1.5 dyes per oligonucleotide were used in these studies.

smFISH imaging

Images were acquired using oil immersion 100× objective on an epifluorescence Olympus BX83 microscope, with an X-Cite 120 PC lamp (EXFO) and an ORCA-R2 digital charge-coupled device (CCD) camera (Hamamatsu) using Cy5 (Cy5-4040C-Zero), Cy3.5 (Cy3.5v1), Cy3 (Cy3-4040C-Zero), and DAPI (DAPI-5060C-Zero) filters (all from Semrock except Cy3.5, from Chroma). Exposure times were 600 ms, 600 ms, 400 ms and 10 ms, respectively. *Z* stacks spanning the entire volume of the cells were acquired by imaging every 300 nm along the *z*-axis. Stage and illumination control of the microscope was achieved using MetaMorph software (Molecular Devices).

Time-lapse microscopy

HSC cultures were maintained as described above using the EVOS Onstage Incubator System on an EVOS FL2 microscope (Thermo Fisher Scientific). Cells were imaged with a 10X objective with phase imaging every 10 min, or were imaged with phase imaging using a 4X objective every 5 min.

Quantification and statistical analysis

All statistical analyses and calculations were made in MATLAB R2018a or MATLAB 2018b except where otherwise noted. All computations were performed on a custom-built PC from AVA with an Intel CORE i7-8700 CPU @3.20 GHz and 32 GB RAM.

Image analysis for smFISH

Detection of single mRNAs was performed by three-dimensional Gaussian fitting of thresholded spots using FISHQUANT (FQ) implemented in MATLAB R2018b. Details on use of FISHQUANT are provided in Supplementary Methods.

Probabilistic transcriptional state assignments

See Supplementary Methods.

Comparison of scRNA-seq and smFISH

See Supplementary Methods.

Summary statistics of mRNA copy number per cell

Extended Data Fig. 3c provides summary statistics for the mRNA counts per cell for *PU.1*, *Gata1* and *Gata2* in primary KL populations. *n* indicates the total number of cells analysed across two separate experiments (CMPs and MEPs) or in a single experiment (GMPs). μ is the arithmetic mean of the number of mRNA molecules per cell, 95% CI is the 95th confidence interval and '%Expressing' is the number of cells with at least 1 detected molecule(s) of mRNA for each gene. All calculations were performed in MATLAB.

Theoretical co-bursting frequencies

Theoretical co-bursting frequencies were calculated by multiplying the probability of a cell having *p* number of transcription sites for gene 1 by the probability of having *q* number of transcription sites for gene 2.

t-Stochastic neighbour embedding maps

t-Stochastic neighbour embedding (*t*SNE) maps of primary KL cells were generated in MATLAB with the 'tsne' function using the mature and nascent mRNA values per cell for each gene as variables.

Transcriptional state assignments: KL

The gating strategy is shown in Extended Data Fig. 5. From all CMPs, large, polyploid megakaryoblasts with hundreds of copies of all three genes are first removed. Next, all cells with *Gata1* >10 are classified as G1/2H (red box, top left histogram). The negative fraction (grey box, top left histogram) is then broken up using *PU.1* and *Gata2*. Given the lack of cleanly separated *PU1* and *Gata2* subpopulations in their respective histograms (top middle and right histograms), the bivariate distribution was used to identify states. P1H (blue box) are identified as *PU.1* > 40 and *Gata2* < 50. G2H (pink box) is identified as *Gata2* > 25, *PU.1* < 40. A small population of $Gata2^{high}PU.1^{high}$ CMPs (yellow box) were difficult to assign. To assess if these were cells destined towards the G1/2H lineage, we compared the inferred transcriptional parameters of the G2H state if we included or excluded these cells from the G2H state. We found only minor changes in transcriptional parameters, with a decrease in k_{on} (the rate of gene activation) and k_{ini} (the rate of RNA polymerase II escape from the paused state) for *PU.1*, and an increase in the k_{off} (the rate of gene inactivation) and k_{ini} for *Gata1* (Supplementary Table 4). However, we also noted that some subset of GMP were $Gata2^{high}PU.1^{high}$. As such, we excluded these cells from all downstream analyses.

Diffusion pseudotime estimation

Diffusion maps based on the mature mRNA counts per CMP cell for *PU.1*, *Gata1* and *Gata2* were generated in MATLAB using the diffusion pseudotime estimation software described in ref. [31]. The diffusion pseudotime maps were generated using a 40-nearest neighbour search with a kernel width of 50. The diffusion map plotted in Fig. 2a shows the first two diffusion components and is coloured according to the transcription state classification scheme described in Extended Data Fig. 5. For the raster spike density plots in Fig. 2b, CMP state subsets were ordered along their inferred pseudotime. For the *Gata* branch, we subsetted on cells in the LES, G2H and G1/2H states. For the *PU.1* branch, we subsetted on cells in the LES and P1H states. Each spike is a cell and the height of the spike is the number of active transcription sites in that cell.

Phase portrait diagrams

Phase portraits are based on similar analyses[39], with the nascent mRNA per cell for a gene given on the *y* axis and the mature mRNA for a gene given on the *x* axis. More than 240 cells were analysed for each gene pair with *PU.1*. Nascent mRNAs are the equivalent number of mature mRNAs found at all active transcription sites for a gene as determined by the integrated intensity of those transcription sites.

Mathematical model of three-gene random telegraph process

All scripts required to run the following model and associated simulations are provided as .m files. Mathematical details on the methods for these sections are found in Supplementary Methods.

Pedigree analysis and kin correlation analysis

Time-lapse movie analysis and mapping to smFISH data. Given the large surface area of the MatTek dish and the need to use two separate microscopes for time-lapse and smFISH imaging, correctly mapping colonies between these systems is exceedingly nontrivial and labour intensive. Imaging the entire surface of the dish for smFISH requires around 500 stage positions with a 100X objective, which is prohibitively long for four-colour acquisition over multiple experiments. As such, we instead realized that the spatial distribution of large megakaryocytes generated during HSC culture creates a reference map

between colonies. These markers can therefore serve as guides during identification of colonies during smFISH imaging acquisition. As such, we used the final frame of the movie to identify regions of the dish in which we could confidently identify colonies on the epifluorescent microscope and imaged these colonies for smFISH. We then manually analysed the time lapse movies for these select colonies in TTT[40]. Single-cell identification within each colony was then performed by manually cross-referencing between the smFISH stacks and the final frame of the movie.

State assignments for HSC. State assignments follow the sequential gating strategy shown in Extended Data Fig. 7, where megakaryocytes are first identified and excluded and then G1/2H cells are identified as cells with more than 10 copies of *Gata1* per cell. P1H macrophage cells are all cells with more than 150 copies of *PU.1* per cell and were similarly excluded from downstream analysis. Of the remaining population, there is no clear threshold that is able to separate G2H cells from P1H or LES (Extended Data Fig. 7b, c, right). As such, we fit both genes to a two-component negative binomial distribution. For the data in Fig. 3, cells were called G2H or P1H rather than LES if they had a probability of assignment to the high-expressing state of *Gata2* or *PU.1*, respectively, of greater than 80%. For the transition dynamics shown in Fig. 4, we used a hard threshold of 75 copies of *PU.1* and used probabilistic assignment for *Gata2*, similar to the treatment of *Esrrb* expression in ref. [37]. An extensive description of this procedure is provided in the Supplementary Methods. This procedure allows for the correction of erroneously assigning a cell in a low *Gata2* state to the high state or vice versa due to the overlap in the negative binomial components.

KCA. A derivation of KCA can be found in ref. [37]. Scripts to perform KCA and consistency checks were adapted from scripts provided by S. Hormoz and M. Elowitz and are provided in the Supplementary Information along with the raw data for all colonies analysed.

In brief, KCA was performed using all colonies analysed across 2 separate experiments for a total of 117 colonies under the assumptions of a stationary, reversible transition matrix between states. Transition probabilities (reported as probability/generation in all figure panels) were inferred at lineage distances of $u = 1$ (sister cells) to $u = 6$ (distant cousin cells). The data in Fig. 4b, c are average inferred transition probabilities for each lineage distance u, and the error bars are the standard error of the mean in those estimates derived by bootstrapping through the data 5,000 times. The script entitled "KCA.m" will generate all the figures found here, and will also save the mean and standard deviation of the inferred transition probabilities between states.

Checking robustness of mRNA cutoff threshold. We used the approach formulated in ref. [37], whereby we re-ran the KCA analysis using different cutoff values for *Gata1* and *PU.1* and then compared these resultant transition matrices to the reference matrix reported in this study (Extended Data Fig. 9).

Checking for spurious state transitions due to partitioning errors. To check our data for spurious transitions inferred during KCA due to asymmetric partitioning of mRNAs, we used two approaches (Extended Data Fig. 10). First, we looked for evidence of such phenomena in our CMP and HSC datasets reported in Fig. 1 and Extended Data Fig. 6. We searched those image banks for sister cells in anaphase–telophase at the time of fixation, separated those cells on the midline, and calculated the correlation coefficient for the mRNA counts for each gene in each population. This analysis revealed very high correlation in the number of mRNAs partitioning to each sister cell.

Second, we used the movies used in the KCA to analyse the correlation in mRNAs between cells having divided within the last hour before fixation at the endpoint. That analysis also revealed considerably high correlation in mRNA values.

Together, these results indicate that our results are probably not substantially affected by partitioning asymmetries of mRNAs during mitosis.

Comparing reversible and irreversible dynamics. To test whether our data were better described by dynamical models containing irreversible transitions (models I and III) compared with those without (models II and IV), we used an approach described in ref. [37]. First, to generate transition matrices for each model, we took the transition matrix derived above (which is model IV) and imposed a new model's dynamics by setting disallowed edges to 0 and re-normalizing each column of the matrix such that all the transition probabilities leaving a state summed to 1.

We then calculated the expected three-state frequencies for $u = 1$ and v (the generational distance of the more distant relative) = 2:4 under each model. We then compared these three state frequencies with the corresponding frequencies for the same values of u and v as derived from the experiment. The data in Fig. 4e (middle) are the average predicted (x axis) and observed (y axis) three-point frequencies. The error bars are for the observed frequencies and derive from bootstrapping through the data 1,000 times. We then calculated the error between the model and observed results as defined by the mean absolute error for all three-point frequencies at a given distance v. The script entitled "ThreePtFreqs.m" found in the associated GitHub page will generate the full analysis reported here.

Calculating time spent in each state. We wrote an algorithm, treeBackTrace, which takes in the structure of a tree together with the final distribution of states among the leaves of this tree, as well as the Markov matrix modelling state transitions between successive generations, and calculates for each leaf node the expected time (measured in number of generations) it spent in each state along its full ancestral trajectory, given the information of the final distribution of states.

To arrive at the conditional expectation, for each possible assignment of states to the intermediate nodes of the tree, one can calculate its probability by multiplying together the resulting transition probabilities indicated by the Markov matrix. For each such assignment and for each leaf node, one can count the distribution of states in its trajectory, and by summing over all such assignments, weighted by their probabilities, and then dividing by the total probability of all such assignments, calculate the conditional expectation mentioned above.

However, such an exhaustive calculation is exponential in time. Instead, we used a divide-and-conquer approach, by breaking up the tree into two subtrees and combining the information from these subtrees, resulting in a linear-time algorithm (see script in Supplementary Data 2).

Calculating the steady state population frequencies. See Supplementary Methods.

Figure generation, plotting and graphics

All figures were generated in MATLAB using either custom written scripts or, for the violin plots in Fig. 1e, the gramm package. Exported .emf or .jpg files were then imported into Adobe Illustrator for cosmetic adjustments such as normalizing the font size across figure panels and adding relevant graphics where needed. Fiji was used to generate jpeg images of all smFISH image stacks. For all images except Extended Data Fig. 4a, we show the filtered image generated during processing in FISHQUANT.

Reporting summary

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

Data availability

All source data used to generate figures are available within the manuscript files or at the GitHub repository (https://github.com/

justincwheat/Single-Molecule-Imaging-of-Transcription-Dynamics-in-Somatic-Stem-Cells) associated with this manuscript. Further information and reasonable requests for resources, reagents and data should be directed to the corresponding author. For data used for generating figures related to kin correlation analysis or simulations (Figs. 2, 4, Extended Data Figs. 8 and 9), separate .mat files have been provided as Supplementary Data 1 and also uploaded to the GitHub repository listed above or are generated upon running the associated scripts. All data are available from the corresponding author upon reasonable request. Source data are provided with this paper.

Code availability

Software written for parameter estimation and stochastic simulations are provided in Supplementary Data 2, (FSP.m, getKLD.m, GSSA.m). Software relevant for Figs. 3 and 4 can also be found in Supplementary Data 2: the code for KCA (KCA.m), generating 3-cell frequency matrices (ThreePtFreqs.m), testing different molecular cutoffs (KCA_thresholdtesting.mlx), and calculating time spent in each state (GenerateAllTrees.m). Data structures for each colony are also provided (Colony[#].mat). All scripts and data files have also been published in a publicly available repository at https://github.com/justincwheat/Single-Molecule-Imaging-of-Transcription-Dynamics-in-Somatic-Stem-Cells. All software generated by other groups used in this study are listed in Supplementary Table 7.

38. Loeffler, D. et al. Mouse and human HSPC immobilization in liquid culture by CD43- or CD44-antibody coating. *Blood* **131**, 1425–1429 (2018).
39. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
40. Hilsenbeck, O. et al. Software tools for single-cell tracking and quantification of cellular and molecular properties. *Nat. Biotechnol*. **34**, 703–706 (2016).

Acknowledgements We thank D. Shechter, K. Gritsman, R. Coleman, J. Biswas, E. Tutucci, M. V. Ugalde and R. Piszczatowski for discussions; F. Mueller for assistance with FISH-QUANT; M. Elowitz and S. Hormoz for the scripts used for KCA; M. Lopez-Jones for assistance in probe design; D. Loeffler and T. Schroeder for input on time-lapse imaging of HSC; and D. Sun for assistance with flow cytometry and cell sorting. R.H.S. is a senior fellow of the Howard Hughes Medical Institute. A.B is an external professor of the Santa Fe Institute. This research was supported by the Ruth L. Kirschstein National Research Service Award F30GM122308-03 and MSTP training grant T32GM007288-43 to J.C.W., U01DA047729 to R.H.S. and R01CA217092 to U.S. U.S. was supported as a Research Scholar of the Leukemia and Lymphoma Society and is the Diane and Arthur B. Belfer Faculty Scholar in Cancer Research of the Albert Einstein College of Medicine. This work was supported through the Albert Einstein Cancer Center core support grant (P30CA013330), and the Stem Cell Isolation and Xenotransplantation Core Facility (NYSTEM grant #C029154) of the Ruth L. and David S. Gottesman Institute for Stem Cell Research and Regenerative Medicine.

Author contributions J.C.W., U.S. and R.H.S. conceptualized the study and designed experiments. J.C.W., A.B. and Y.S. conceptualized mathematical models. J.C.W. performed all experiments and generated all data in the manuscript. J.C.W. performed the mRNA analyses, transcriptional parameter fitting, stochastic simulations, scRNA-seq analyses, and kinship analyses. M.W. provided essential scripts for scRNA-seq analyses. Y.S. and A.B. developed the analyses related to the history of state transitions conditional on pedigree structure. J.C.W. wrote the manuscript and generated all figures and data visualizations. J.C.W., U.S., R.H.S., A.I.S., Y.S., A.B. and M.W. reviewed and edited the manuscript.

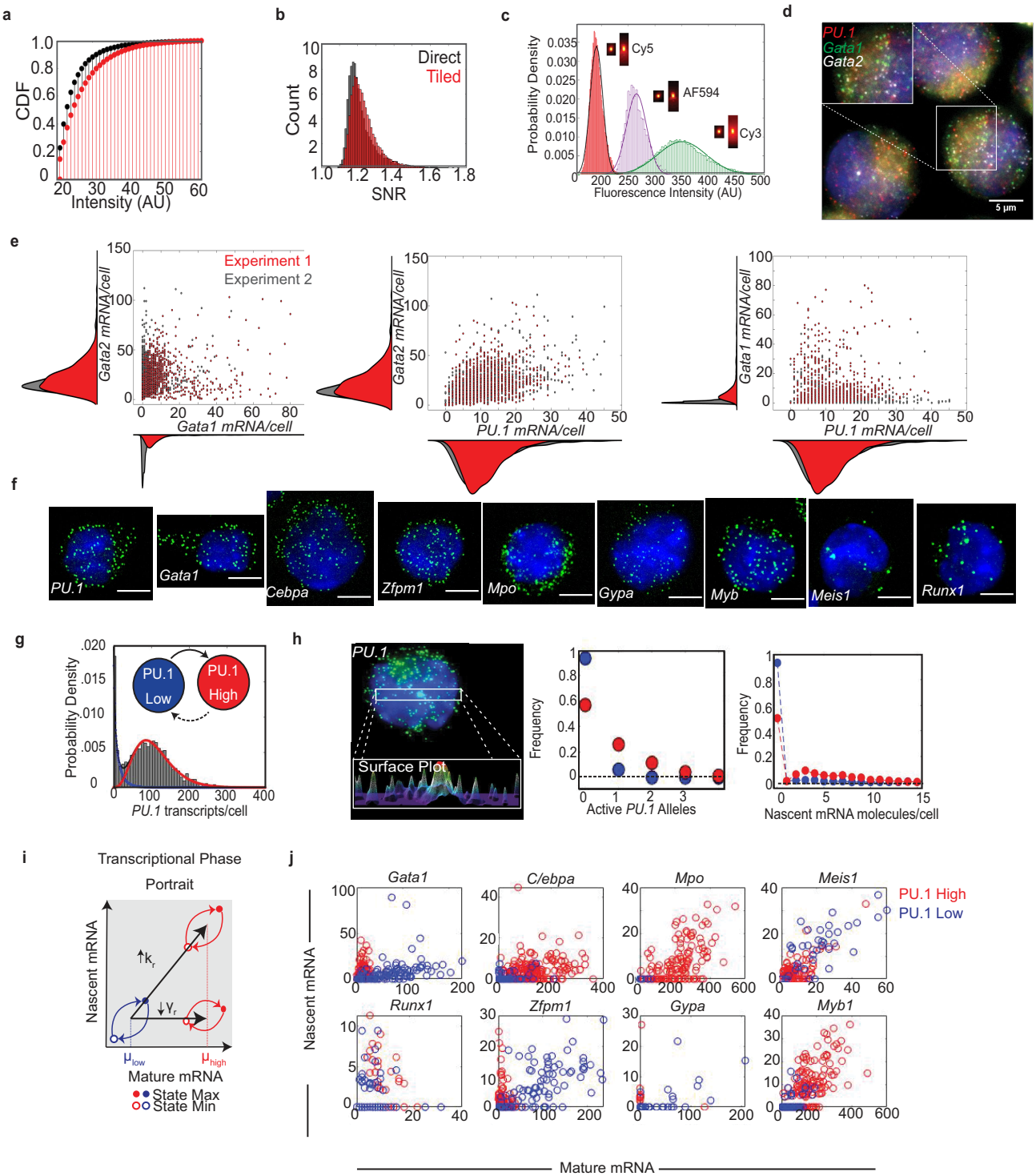
Competing interests The authors declare no competing interests.

Additional information

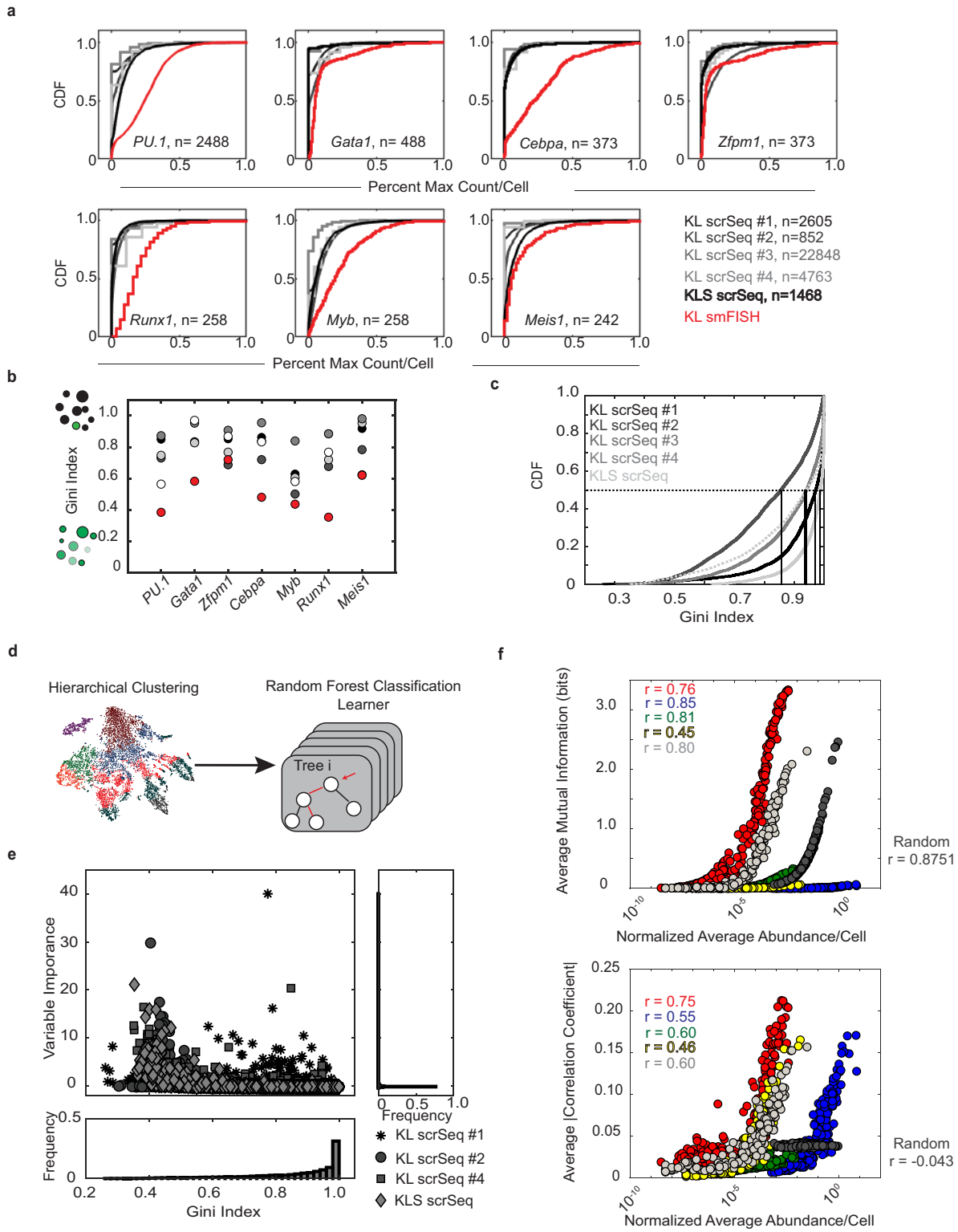
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2432-4.

Correspondence and requests for materials should be addressed to U.S.

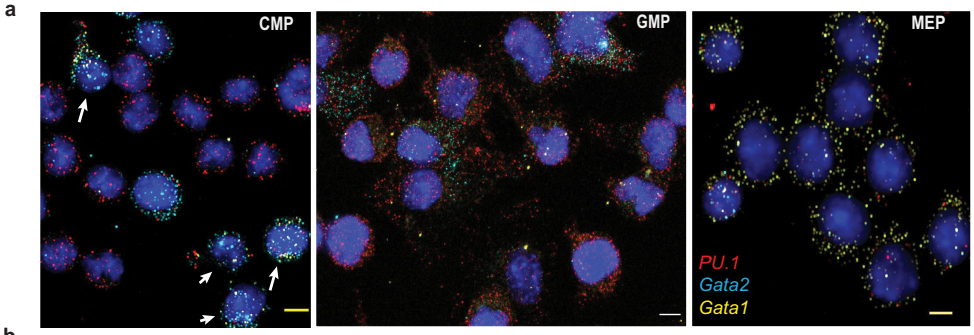
Peer review information *Nature* thanks Thomas Gregor, Ellen Rothenberg and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

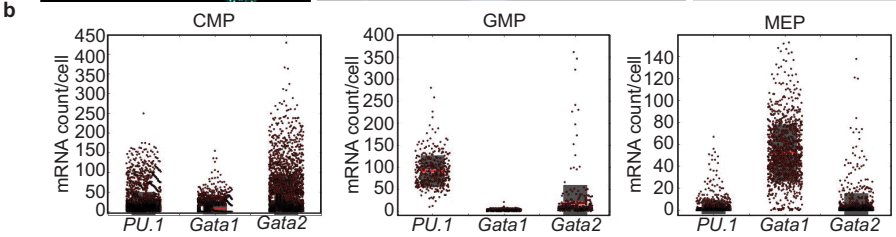


Extended Data Fig. 1 | Transcriptional dynamics of genes conditional on *PU.1* state. a, **b**, Cumulative distribution function (CDF) of spot intensity (**a**) and histogram of signal-to-noise ratio (SNR) of spot intensity to local background intensity (**b**) are shown for all spots that passed intensity and 3D point-spread function (PSF) fit thresholding in FISH-QUANT. **c**, Probability densities for fluorescence (corresponding to mRNA molecules) in HPC-7 cells for Cy3-, Alexa Fluor 594- and Cy5-labelled readout probes. Insets are *XY* and *XZ* average PSFs for each fluorophore. The overlaid line is the fit to a Gaussian distribution. More than 10,000 spots were obtained per fluorophore. **d**, Representative images three-colour smFISH for *PU.1* (Cy5, red), *Gata2* (Cy3, white) and *Gata1* (AF594, green) in HPC-7 cells. Scale bar, 5 µm. **e**, Bivariate distributions of *Gata1* and *Gata2* (left), *Gata2* and *PU.1* (middle) and *PU.1* and *Gata1* (right) in two independent experiments ($n > 400$ cells per experiment) with HPC-7 cells. **f**, Representative images of multiplexed smFISH between *PU.1* and eight other haematopoietic genes in Kit$^+$Lin$^-$ bone marrow from wild-type mice ($n = 258–2,488$ cells for each gene, derived from a single experiment; scale bar, 5 µm). **g**, Probability distribution for *PU.1* mRNA per cell in KL cells from bone marrow from wild-type mice. Overlaid are the high (red) and low (blue) components of the two-component negative binomial distribution fitted to the data. **h**, Comparison of PU.1 bursting kinetics between high and low states. Left, representative images from smFISH for *PU.1* with a single, large transcription site in the nucleus. Middle, frequency of cells with the indicated number of active *PU.1* transcription sites. Right, frequency distribution of summed nascent mRNA per cell in each *PU.1* state. **i**, Schematic demonstrating a hypothetical transcriptional phase portrait. **j**, Phase portraits for each gene based on the *PU.1* state of the cell.



Extended Data Fig. 2 | Comparative analysis of smFISH and scRNA-seq. **a**, CDF plots of mRNA per cell for five scRNA-seq datasets and smFISH. Data are normalized to the maximum count for each gene in each dataset. **b**, Calculated Gini index for seven transcription factor mRNAs in each scRNA-seq dataset (white through to black) and smFISH (red). **c**, CDF plots of Gini indices for all five scRNA-seq datasets (See Supplementary Table 2 for gene list). **d**, Schematic of hierarchical clustering followed by random forest classification to identify important variables for cluster assignment. **e**, Variable importance plotted against Gini index for four scRNA-seq datasets. The bottom and right panels show marginal distributions of Gini index and variable importance, respectively. **f**, Plot of average mutual information (top) or average absolute value of the Pearson's correlation coefficient (bottom) versus normalized abundance of $n = 200$ randomly selected genes against all other genes in the dataset. The r values listed are the correlation coefficients. See Supplementary Discussion for further details on the analyses performed.



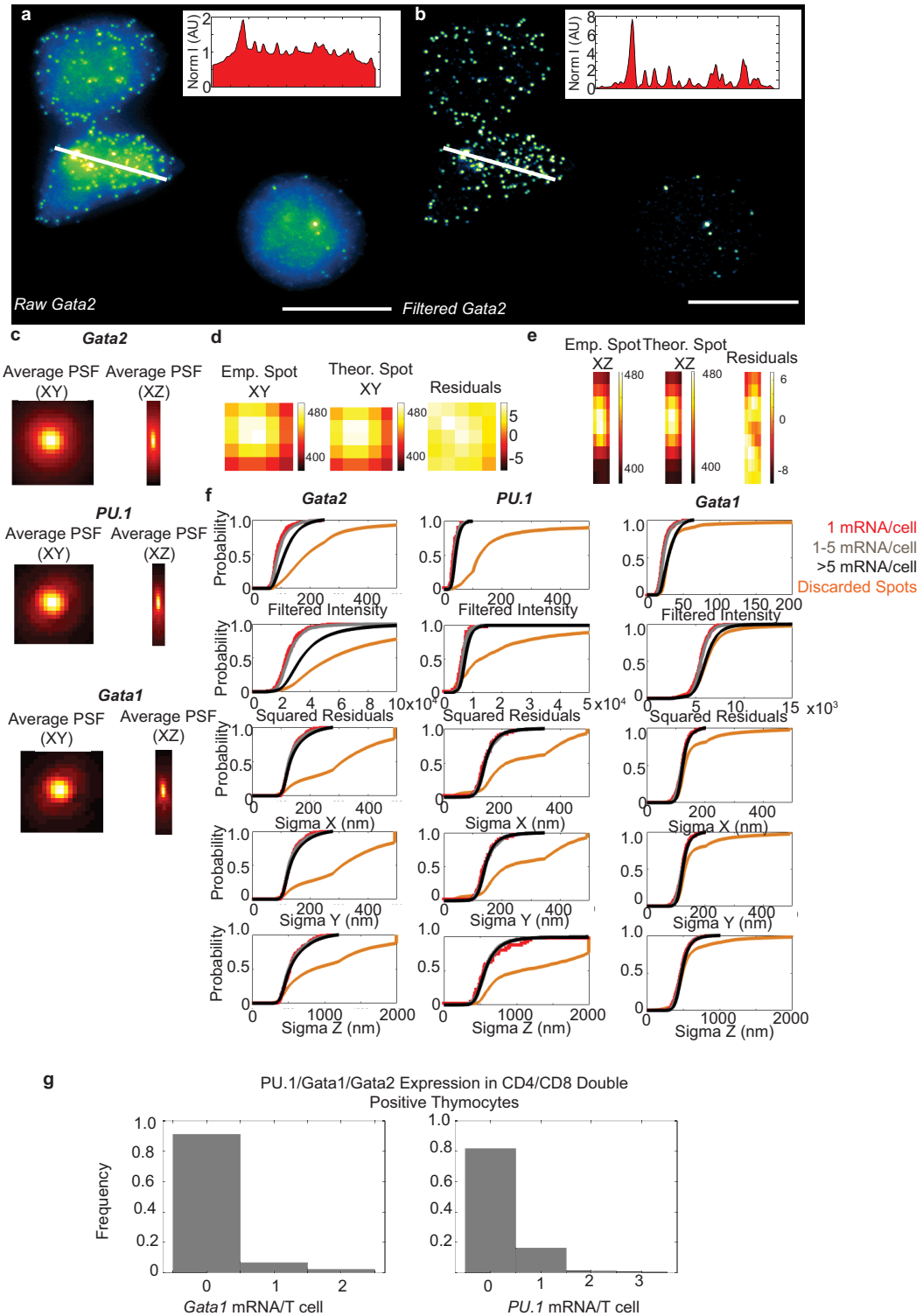


c

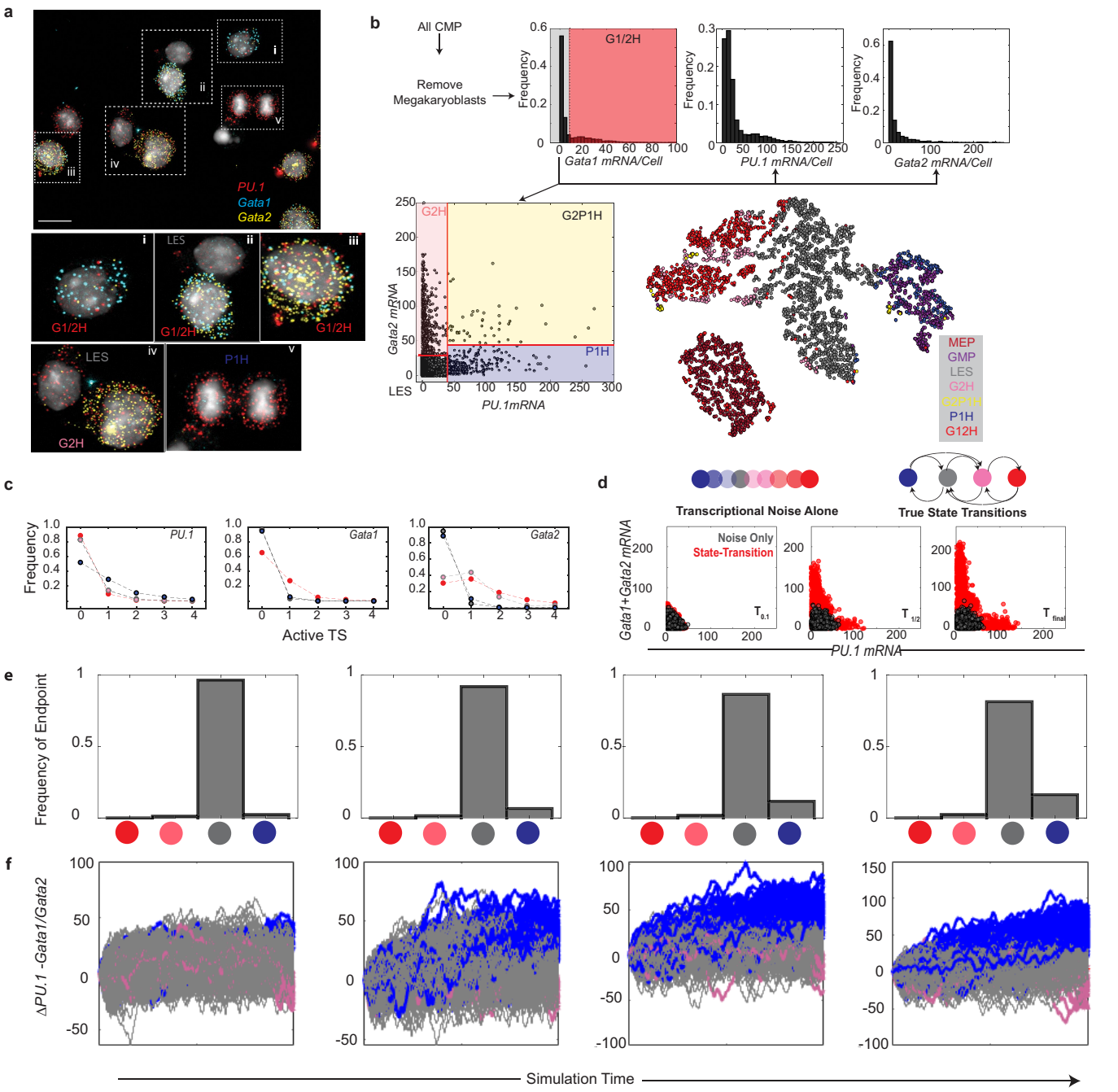
Table 1: Statistics for Primary KL Populations

	CMP (N=3174)			GMP (N=364)			MEP (N=1113)		
	μ (95% CI)	range	%Expressing	μ (95% CI)	range	%Expressing	μ (95% CI)	range	%Expressing
PU.1	21 (20-22)	0-250	97%	91 (87-95)	0-270	97%	3 (2-4)	0-67	68%
Gata1	8 (7-9)	0-155	64%	3 (2-3)	0-21	90%	52 (51-54)	0-153	99%
Gata2	42 (40-44)	0-430	96%	18 (13-22)	0-361	99%	4 (3-5)	0-138	68%

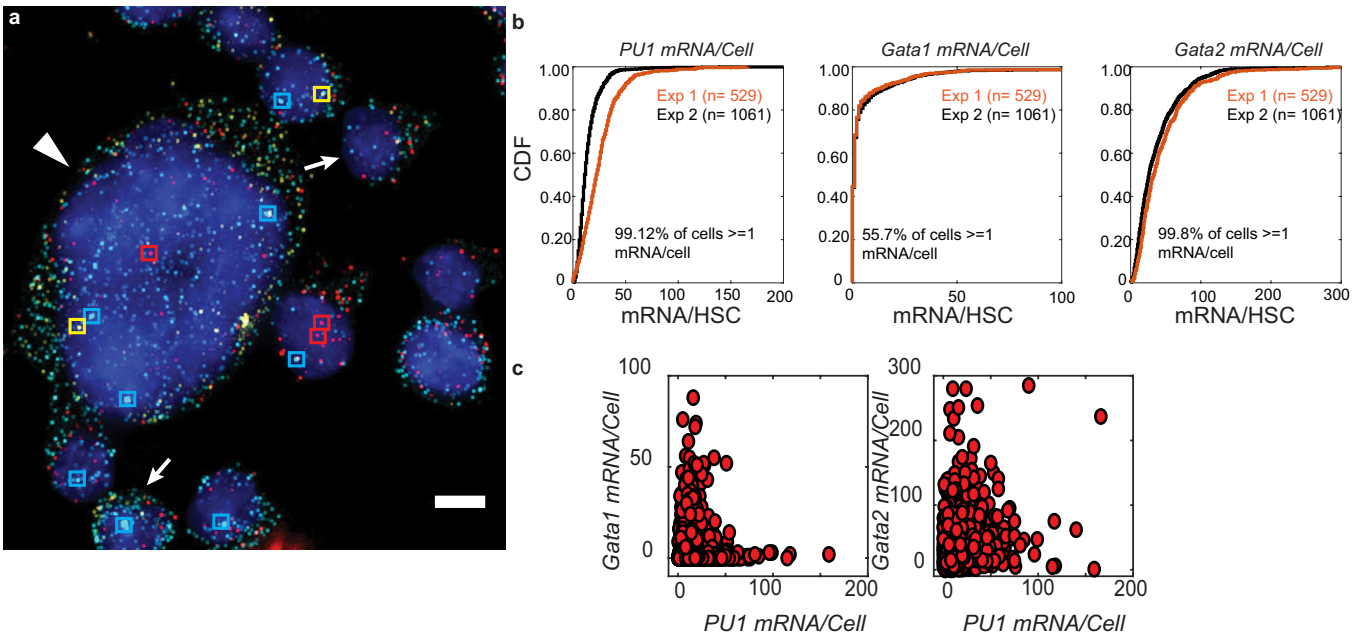
Extended Data Fig. 3 | Summary statistics of mRNA copy number for primary KL. a, Representative images of CMPs, GMPs and MEPs stained by smFISH for *PU.1*, *Gata1* and *Gata2*. Scale bars, 5 μm. Arrows point to CMPs co-expressing all three mRNAs. **b**, Boxplots of mRNA count per cell, overlaid with single-cell mRNA values (dots). The pink box is the 95% confidence interval, the red line is the mean expression, the grey box is ±s.e.m. **c**, Table of summary statistics for each gene. Data for **a**–**c** are derived from two experiments (CMPs and MEPs) or a single experiment (GMPs). The sample size is listed in **c**.



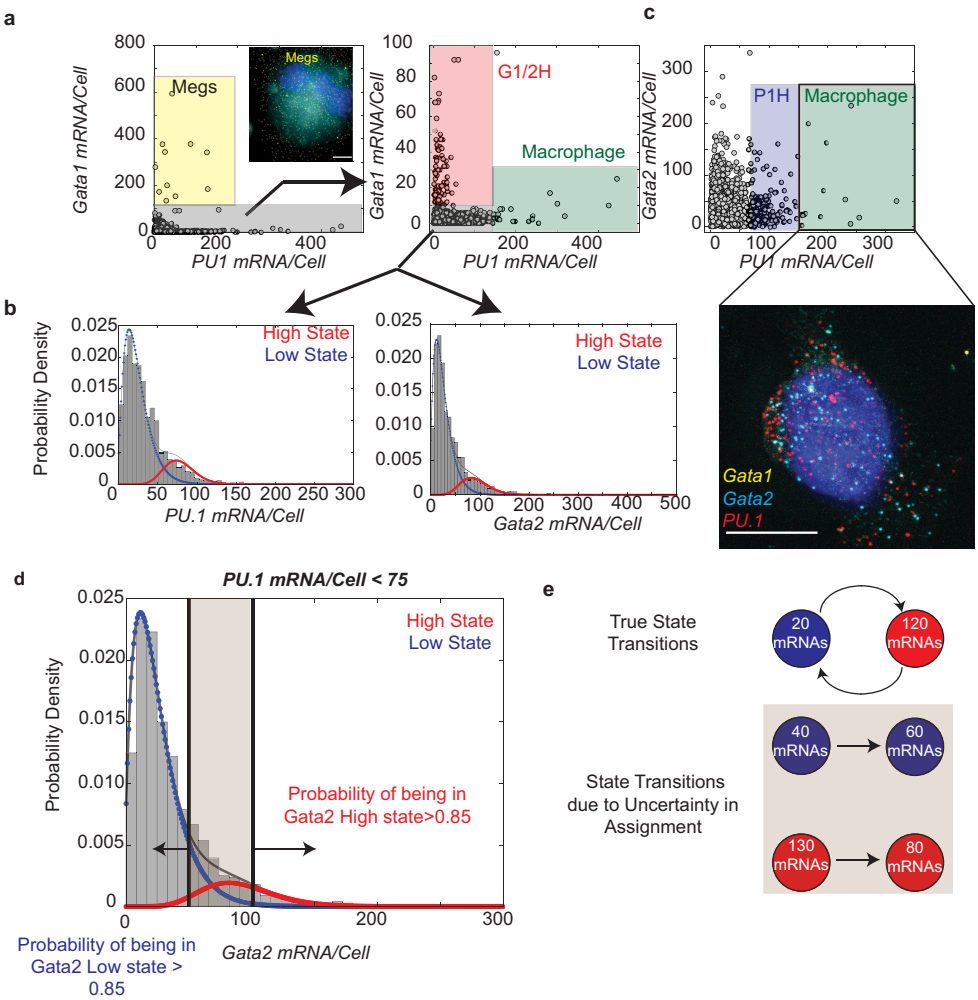
Extended Data Fig. 4 | Spot detection in FISH-QUANT and spot calling in T lymphocytes. a, **b**, Comparison of raw (**a**) and filtered (**b**) smFISH images from CMPs (representative of more than 2 experiments in CMPs; spot quality is consistent with all reported experiments in this manuscript). The insets show line intensity plots; the white line on the cells indicates from where the plots were obtained. Scale bars, 10 µm. **c**, Average PSF in *XY* (left columns) and *XZ* (right columns) for each gene from all detected spots from the CMPs dataset. **d**, **e**, Empirical (left) versus theoretical (middle) PSF and residuals (right) in the *XY* (**d**) and *XZ* (**e**) planes. **f**, CDFs for all spots passing the initial intensity thresholding for filtered intensity (top row), squared residuals (second row) and width of spots in *X*, *Y*, and *Z* in nanometres (third to fifth rows, respectively). Spots are separated on the basis of those arising from cells with more than five copies of mRNA per cell, between two and five copies per cell, and one copy per cell. Discarded spots that failed 3D fitting are shown in orange. **g**, mRNA detection in primary $CD4^+CD8^+$ thymocytes ($n = 136$ for *Gata1*, $n = 154$ for *PU.1*).



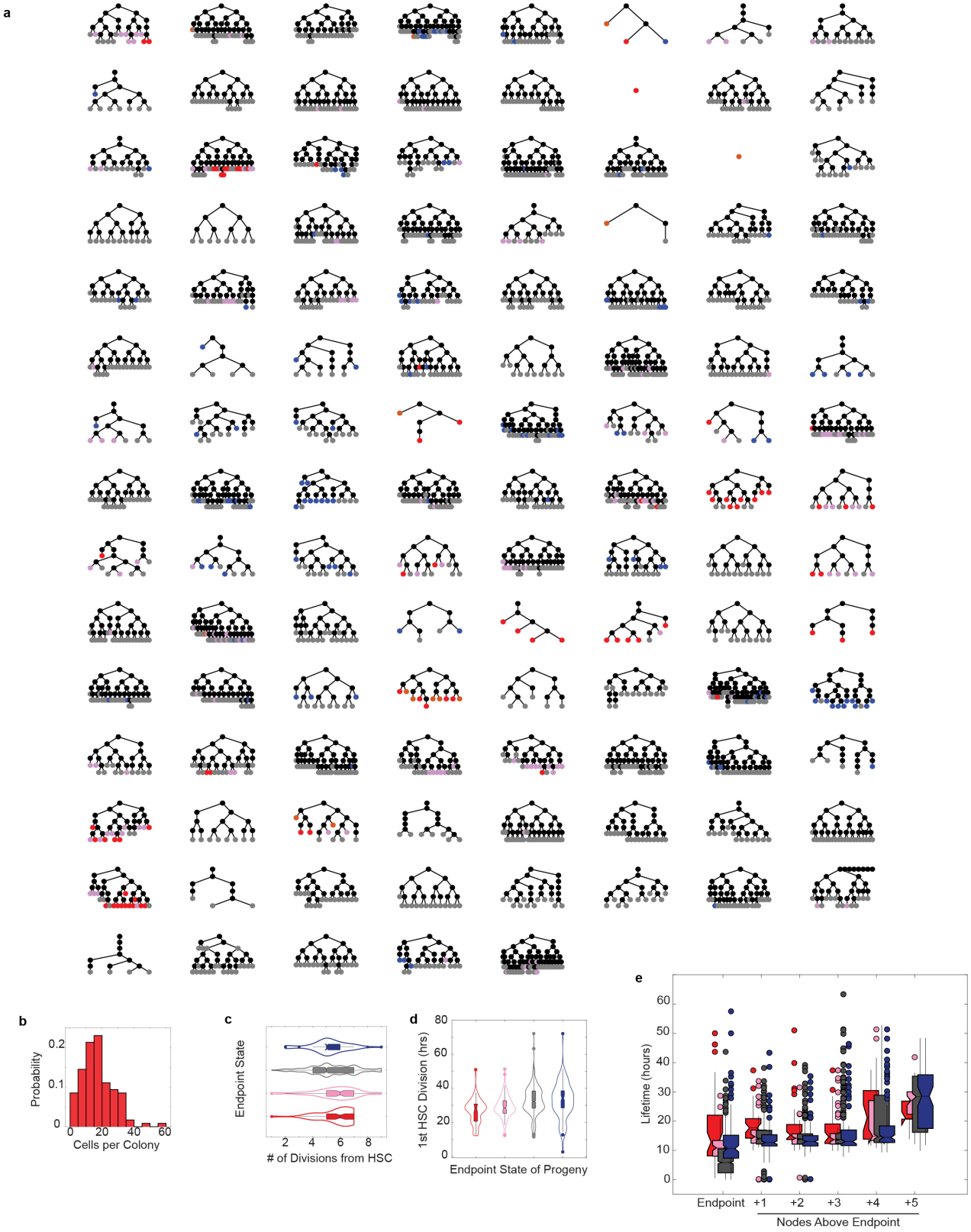
Extended Data Fig. 5 | Gating strategy to assign CMP to states. **a**, Representative images of CMPs in different states. Scale bar, 10 µm. **b**, Gating scheme for assigning CMPs to transcriptional states. See Supplementary Discussion for details on the gating strategy. The t-SNE plot demonstrates the proximity of states to one another and to immunophenotypic GMPs and MEPs. Images and analyses derived from experimental datasets reported in Fig. 1 and Extended Data Fig. 3. **c**, Frequency distribution of transcriptional bursting for each gene in each transcriptional state. The *x* axis is the number of active alleles. **d**, Top, schematic of 'states' being the consequence of simple transcriptional noise of the LES state (right) versus truly separate transcriptional states (right) that require transition events (arrows). Bottom, time-dependent behaviour of simulated cells in a noise only (grey) or state transition system (red) shown as a bivariate plot of *Gata1* + *Gata2* copy number against *PU.1* copy number. T indicates the elapsed simulation time as a fraction of the final time. **e**, **f**, Gillespie simulations of state transitions, modulating half-life alone. If a transition to another state occurs by noise alone, the cell changes the mRNA half-life of only the mRNA defining that state. **e**, **f**, Endpoint states reached in the simulations ($n = 10{,}000$) (**e**) and 1,000 representative simulation trajectories (**f**), colour-coded on the final endpoint state. Each panel is a different factor change in the mRNA half-life, with the far-left panel as the reference (that is, the half-lives used in Fig. 2), and the other panels showing 2× (second from left), 3× (second from right), and 4× (far-right).



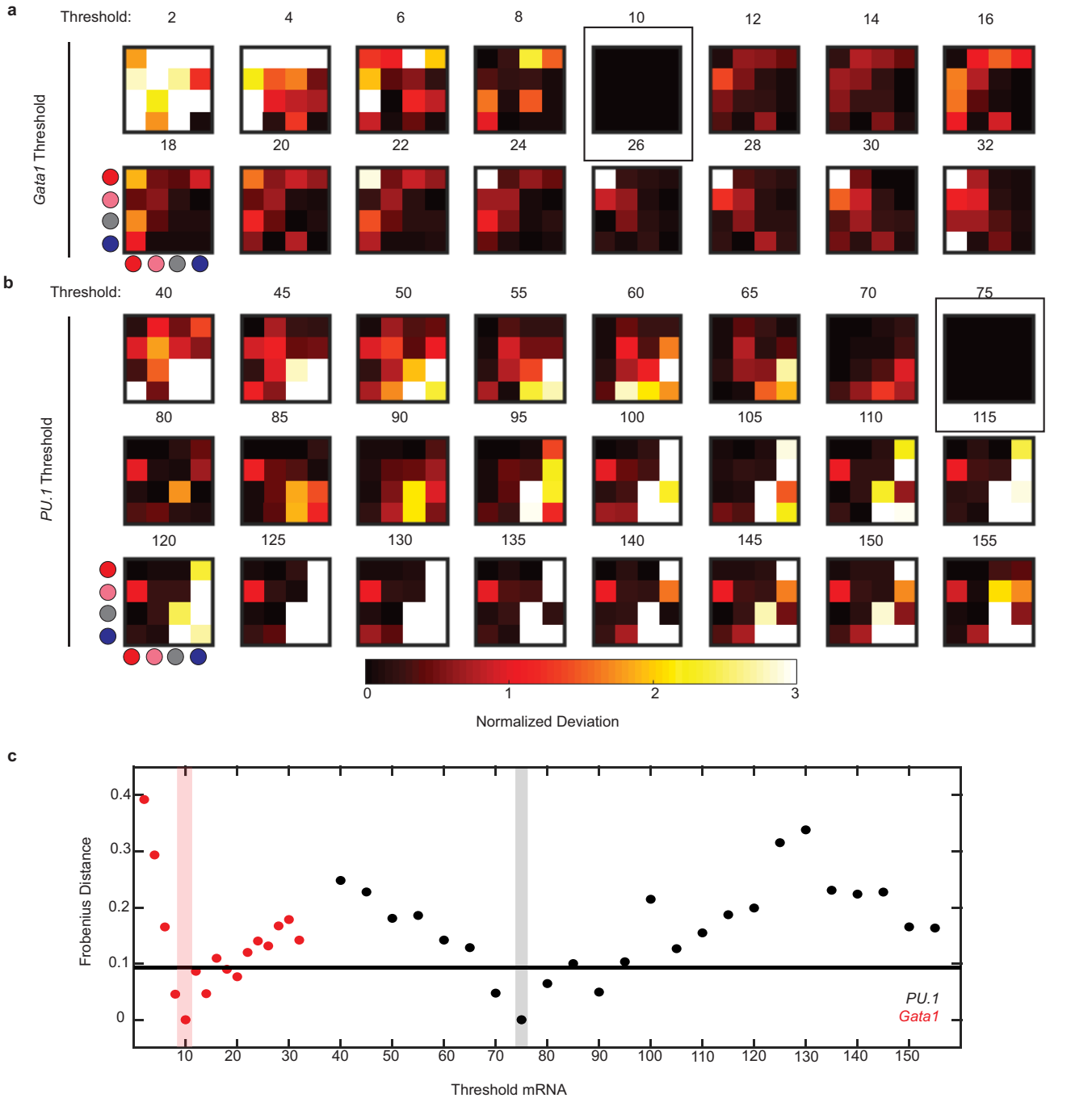
Extended Data Fig. 6 | Seventy-two-hour progeny of HSCs. a, Representative images of HSC progeny. *PU.1*, red; *Gata2*, cyan; *Gata1*, yellow. Transcription sites are demarcated with boxes. Full arrows indicate triple-positive cells, and the arrowhead marks a megakaryocyte. Representative images from two separate experiments. **b**, CDFs for mRNA counts per HSC progeny. The number of cells with greater than or equal to 1 mRNA per cell is indicated. Two separate experiments, with *n* values indicated on the graphs. **c**, Bivariate distributions of *PU.1* versus *Gata1* (left) and *PU.1* versus *Gata2* (right).



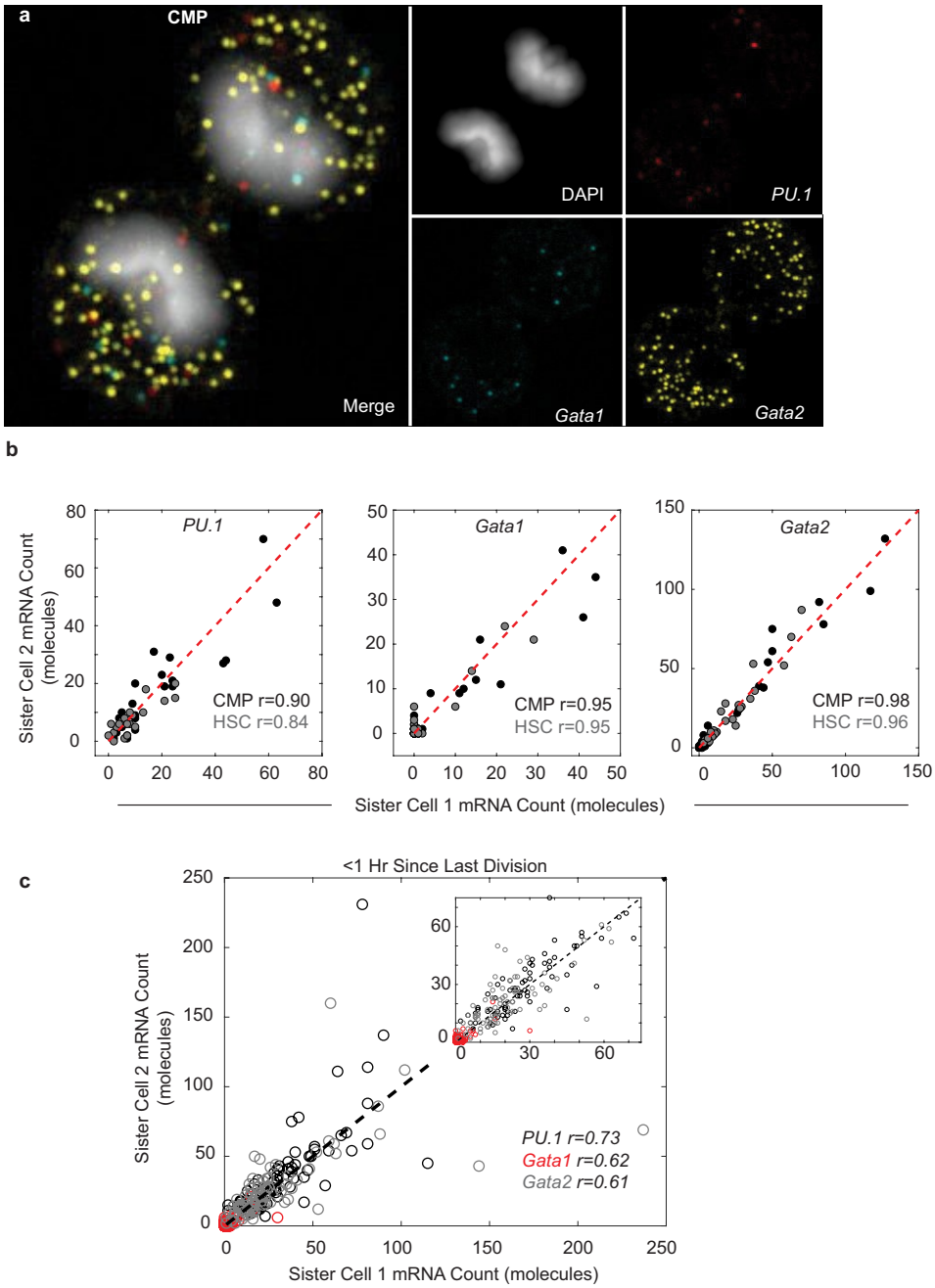
Extended Data Fig. 7 | State assignments for HSC progeny. a, Gating strategy. Left, removal of megakaryocytes occurs first. Right, cells with more than 10 copies of *Gata1* are assigned to G1/2H, whereas cells with more than 150 copies of *PU.1* are assigned to macrophage. **b**, Probability density distributions for *PU.1* (left) and *Gata2* (right) with overlaid fits for a two-component negative binomial distribution amongst cells after removing megakaryocytes, G1/2H, and macrophage . **c**, Bivariate distribution of the same cells. Contrary to the case in CMPs, the population of *Gata2*^{high}*PU.1*^{high} HSC progeny all had morphological characteristics similar to macrophage-like cells seen in GMP datasets, which also were *Gata2*^{high}*PU.1*^{high} (see Extended Data Fig. 3). As such, all cells for which *PU.1* > 75 and < 150 were assigned to P1H. **d**, Probability distribution for *Gata2* in the remaining cells, fit with a two-component negative binomial. **e**, A distribution such as that in **d** cannot be definitively separated into high and low components owing to overlap in the distributions; therefore, cells are assigned probabilistically during KCA to the G2H or LES state in order to correct for false transitions arising from uncertainty in the assignment. See Supplementary Discussion for more details on the rationale and implementation of probabilistic gating.



Extended Data Fig. 8 | HSC colony data. a, Endpoint cells are the leaves on each pedigree. Note that edge lengths are not scaled on time between divisions, and all endpoint cells are 96 h from the start of the experiment. Cells are colour-coded according to the colour scheme used throughout the manuscript. Megakaryocytes are labelled in orange. Nodes (cells) observed upstream of the endpoint (that is, no transcriptional data are available) are coloured black. **b**, Histogram of number of progeny from a single HSC. **c**–**e**, Proliferation phenotypes of cells based on endpoint state identity (P1H, $n = 137$; LES, $n = 1,571$; G1/2H, $n = 81$; G2H, $n = 166$). Cell lifetimes in **e** are the time interval between cell birth (last division) and the next cell division or cell death. Violin plots are normalized to area, with the centre box-and-whisker plots showing the mean (line), standard deviation (box) and 95% confidence interval (whiskers). In **e**, single dots represent outliers in the 99th percentile.



Extended Data Fig. 9 | Robustness of inferred transition matrix to mRNA threshold. a, Normalized deviation in the inferred transition matrices for each indicated threshold ($n = 200$ bootstrapping iterations) of *Gata1* mRNA per cell relative to the reference matrix reported in this manuscript (cutoff = 10 mRNA per cell). The reference matrix is boxed. For any given transition (that is, matrix entry), the initial states are the columns, final states are rows. The colour code is the same as is used elsewhere in the manuscript. **b**, As in **a** for *PU.1* (cutoff in manuscript = 75 mRNA per cell). **c**, Frobenius distance $\sqrt{\sum_{ij}(T_{i,j_{\mathrm{ref}}} - T_{i,j_{\mathrm{test}}})^2}$ between each matrix versus the reference transition matrix. The solid black line indicates the background Frobenius distance derived from statistical uncertainty in the reference transition matrix, derived by bootstrapping through the analysis $n = 1{,}000$ times and picking random transition rates from a Gaussian distribution defined by inferred mean and standard deviation of the transition matrix. Frobenius distance values above this line significantly differ from the matrix reported in the manuscript.



Extended Data Fig. 10 | Analysis of mRNA partitioning errors. **a**, Representative image of a CMP in late anaphase. **b**, mRNA copy number in each sister cell in CMPs ($n = 52$) and HSCs ($n = 46$). r is the Pearson's correlation coefficient for sister-cell mRNA copy number; the red dashed line is $y = x$. **c**, Correlation of mRNA levels between HSCs that divided within the last 1 h ($n = 171$). Pearson's correlation coefficients (r) for each gene are listed.

nature research

Corresponding author(s): Ulrich Steidl M.D. Ph.D.

Last updated by author(s): 03/10/2020

Reporting Summary

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

Statistics

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

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data of flow cytometry was collected with FACSDiva 8 or Summit v62 for the FACS experiments on BD FACSAria II system or MoFlo Astrios EQ system, respectively. Microscope controlled with Metamorph (Molecular Devices, Inc.)
Data analysis	1. Spot calling in smFISH datasets: FISHQuant, version 3 (Mueller et al, 2012). 2. Kin Correlation Analysis (Hormoz et al, 2016). 3. Calculation of Entropy: Entropy Package (Hausser and Strimmer, 2009) 4. Variable Importance in Random Forest Classifier: varSelRF (Diaz-Uriarte and de Andres, 2006). 5. Clustering of scrSeq: DynamicTreeCut (Langfelder and Zhang, 2016) 6. Diffusion Pseudotime analysis (Haghverdi et al, 2016) 7. MATLAB, 2017-2019 version 9.2..0 (R2017a) through version 9.6 (R2019a), Natick, Massachusetts: The MathWorks Inc. 8. R Studio (1.1.456) 2018 (RStudio: Integrated Development for R. RStudio, Inc, Boston,MA) 9. Fiji (Schindelin et al, 2012) 10. Analysis specific to this paper is deposited at: https://github.com/justincwheat/Single-Molecule-Imaging-of-Transcription-Dynamics-in-Somatic-Stem-Cells

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

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

All source data used to generate figures are available within the manuscript files or at the Github repository (https://github.com/justincwheat/Single-Molecule-Imaging-of-Transcription-Dynamics-in-Somatic-Stem-Cells) associated with this manuscript. Further information and reasonable requests for resources, reagents and data should be directed to and will be fulfilled by the Lead Contact Ulrich Steidl (ulrich.steidl@einstein.yu.edu). All raw data used for the generation of figures has been added as Source data. For data used for generating figures related to kin correlation analysis or simulations (Figures 2, 4, Extended Data 8 and 9), separate .mat files have been uploaded to the Github repository listed above or are generated upon running the associated scripts

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

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

Life sciences study design

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

Sample size	Given our experience in pilot experiments whereby we discovered that PU.1 and Gata1 were infrequently bursting in CMP (~10-20% of cells with an active site), we therefore decided to image at least 2500 cells in order to observe at least 250 transcription sites for transcriptional parameter fitting. For KCA, we relied on prior related work that analyzed similar numbers of colonies (Hoppe et al, Nature 2016).
Data exclusions	For smFISH analysis, all cells with nuclei that were clearly blebbed and apoptotic were excluded from all smFISH analysis. Also, cells were excluded if we could not identify clear borders between neighboring cells. For the time lapse microscopy experiments, only colonies which could be mapped successfully between the smFISH images and the movie were included in the analysis of transcriptional states. Megakaryocytes, as defined by nuclear size and DAPI intensity >3x mean of data set or cells which underwent endomitosis during time lapse microscopy, were also removed as they cannot by definition by analyzed with KCA.
Replication	The data used in Figures 3,5, and 6 were repeated twice. For the comparisons with scrSeq, we performed a single experiment analyzing >200 cells per smFISH reaction.
Randomization	Randomization is not relevant to this study as it is an observational study of transcript counts in primary HSPC
Blinding	Blinding is not possible in this study owing to the nature of the experimental studies but was deemed unnecessary in this observational study.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Anti-Mouse CKIT (clone 2B8) APC (1:250) Biolegend Cat# 105811 Anti-mouse Sca1 (clone D7) APC Cy7 (1:250) Biolgened Cat# 108126 Anti-Mouse CD34 (clone RAM34) FITC (1:100) eBioscience Cat# 553733 Anti-mouse CD150 (clone TC15-12F12.2) (1:250) PE Biolegend Cat# 115904

Anti-Mouse CD16/32 (clone 93) PE-Cy7 (1:500) Biolegend Cat# 101318
Anti-Mouse CD48 (clone HM48-1) BV421 (1:250) Biolegend Cat# 103428
Anti-Mouse B220 (clone RA3-6B2) Biotin (1:1000) Biolegend Cat# 103204
Anti-Mouse Gr1 (clone RB6-8C5) Biotin (1:1000) Biolegend Cat# 108404
Anti-Mouse CD11b (clone M1/70) Biotin (1:1000) Biolegend Cat# 101204
Anti-Mouse Ter119 (clone Ter119) Biotin (1:1000) Biolegend Cat# 116204
Anti-Mouse CD127 (clone A7R34) Biotin (1:1000) Biolegend Cat# 135006
Anti-Mouse CD19 (clone 6D5) Biotin (1:1000) Biolegend Cat# 115504
Anti-Mouse CD3 (clone 17A2) Biotin (1:1000) Biolegend Cat# 100244
Anti-Mouse CD4 (clone RM4-5) Biotin (1:1000) Biolegend Cat# 100508
Anti-Mouse CD8 (clone 53-6.7) Biotin (1:1000) Biolegend Cat# 100704
Streptavidin Pacific Orange (1:1000) Thermo Cat# S32365
Anti-Mouse CD43 (eBioR2/60) Biotin (10ug/mL) Thermo Cat# 13-0431-82

Validation

Standard FACS antibodies obtained from widely used commercial providers were used in this study. All antibodies were validated through positive and negative controls, as well as isotype control antibodies.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

HPC-7 cells were obtained from Dr. Omar Abdel-Wahab.

Authentication

No authentication was performed.

Mycoplasma contamination

Cell lines were not tested for mycoplasma.

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified cell lines were used in this study.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

6-10 week old mus musculus, C57/Bl6, both male and female.

Wild animals

This study does not involve wild animals

Field-collected samples

This study does not involve field collected samples.

Ethics oversight

All experiments were approved by the Institutional Animal Care and Use Committee of the Albert Einstein College of Medicine Institute (2016-1003). All procedures were performed in accordance with guidelines from the Institutional Animal Care and Use Committee of the Albert Einstein College of Medicine Institute (2016-1003).

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

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Bone marrow was obtained by crushing femurs, tibiae, pelvis, and vertebrae with a mortar and pestle in MACS buffer (PBS, 1% FBS, 1mM EDTA). Samples were then filtered and subjected to density centrifugation to remove granulocytes and erythrocytes on Ficoll. After removing the buffy coat, samples were rinsed twice prior to staining with FACS antibody cocktails on ice and protected from light.

Instrument

BD FACSAria II system or MoFlo Astrios EQ system

Software

FACSDiva 8 or Summit v62

Cell population abundance

Population purity was checked by a post sort analysis to ensure >98% target population.

Gating strategy | Gating strategy is shown in Supplemental Figure 1 and detailed in the text

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