Published in final edited form as:

Nature. 2020 July ; 583(7816): 431–436. doi:10.1038/s41586-020-2432-4.

Single Molecule Imaging of Transcription Dynamics in Somatic Stem Cells

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,*]

[1]Department of Cell Biology, Albert Einstein College of Medicine, Bronx, New York, 10461, USA

[2]Ruth L. and David S. Gottesman Institute for Stem Cell Research and Regenerative Medicine, Albert Einstein College of Medicine, Bronx, NY, 10461, USA

[3]Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, New York, United States of America

[4]Dominick P. Purpura Department of Neuroscience, Albert Einstein College of Medicine, Bronx, New York, United States of America.

[5]Department of Pathology, Albert Einstein College of Medicine, Bronx, New York, United States of America.

[6]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico, United States of America.

[7]Department of Anatomy and Structural Biology, Albert Einstein College of Medicine, Bronx, New York, 10461, USA

[8]Gruss-Lipper Biophotonics Center, Albert Einstein College of Medicine, Bronx, NY, 10461, USA.

[9]Janelia Research Campus of the HHMI, Ashburn, VA, 10461, USA

[10]Department of Medicine (Oncology), Albert Einstein College of Medicine-Montefiore Medical Center, Bronx, NY, 10461, USA.

[11]Albert Einstein Cancer Center, Albert Einstein College of Medicine, Bronx, NY, 10461, USA.

Summary

Molecular noise is a natural phenomenon inherent to all biological systems[1,2]. How stochastic processes give rise to the robust outcomes supportive of tissue homeostasis is a conundrum. Here, to quantitatively investigate this issue, we use single-molecule mRNA FISH (smFISH) on

[*]**Correspondence and request for materials** should be addressed to US, ulrich.steidl@einstein.yu.edu.

Author contributions. JCW, US, and RHS conceptualized the study and designed experiments. JCW, AB, and YS conceptualized mathematical models. JCW performed all experiments and generated all data in the manuscript. JCW performed the analysis of mRNA analyses, transcriptional parameter fitting, stochastic simulations, scRNAseq analyses, and kinship analyses. MW provided essential scripts for scRNAseq analyses. YS and AB developed the analyses related to the history of state transitions conditional on pedigree structure. JCW wrote the manuscript and generated all figures and data visualizations. JCW, US, RHS, AS, YS, AB, AS, MW reviewed and edited the manuscript.

Competing interests. The authors declare no competing interests.

Supplementary information is available for this paper

stem cells derived from hematopoietic tissue to measure the transcription dynamics of three key transcription factor (TF) genes: *PU.1, Gata1* and *Gata2*. Our results indicate that infrequent, stochastic bursts of transcription result in the co-expression of these antagonistic TF in the majority of hematopoietic stem and progenitor cells. Moreover, by pairing smFISH to time-lapse microscopy and the analysis of pedigrees, we find that while individual stem cell clones produce offspring that are in transcriptionally related states, akin to a transcriptional priming phenomenon, the underlying transition dynamics between states are nevertheless best captured by stochastic and reversible models. As such, the outcome of a stochastic process can produce cellular behaviors that may be incorrectly inferred to have arisen from deterministic dynamics. In light of our findings, we propose a model whereby the intrinsic stochasticity of gene expression facilitates, rather than impedes, 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]. Due to low copy number fluctuations[3–5], spatial and temporal partitioning of reactions[6], and the hard, physical bounds limiting 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 robust production of mature cell types derive from these intrinsically stochastic processes.

Hematopoiesis is a paradigmatic stem cell differentiation model rooted in the hematopoietic stem cell (HSC; Fig. 1a). Single-cell RNA sequencing (scRNAseq) studies have suggested that the gene expression states underlying terminal branches of the hematopoietic tree arise early and continuously during a multi-step differentiation process through populations of increasingly restricted progenitor cells[8,9]. Transcription factors (TF) are thought to play 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 TF gene expression is fundamental.

Single molecule imaging in primary HSPC

Quantitatively evaluating transcriptional noise requires single-cell techniques with molecular sensitivity. As such, we adapted a single-molecule FISH (smFISH) technique, the gold standard for single-cell mRNA analysis, to study this problem in primary hematopoietic stem and progenitor cells (HSPC)[10,11]. Owing to the short mRNA length and high GC content of some critical hematopoietic TF, we opted for a two-step hybridization strategy to increase signal to noise (Fig 1b, Supplementary Methods 1)[12,13]. We first tested this technique on *PU.1* (*Sfpi1, Spi1*), a TF with essential activating functions in myeloid cell development[14,15], a negative repressor of erythroid differentiation[16,17], and whose expression is deregulated during leukemogenesis[18–20]. Two step smFISH markedly increased spot intensity, signal-to-noise, and increased the number of detectable *PU.1* mRNA/cell compared to 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).

As smFISH also allows for the direct observation of active transcription sites[21], we first asked how genes predicted to be co-regulated with *PU.1* correlated in both mature mRNA counts and transcriptional activity. We performed multiplexed smFISH for *PU.1* and 8 critical hematopoietic genes (TF Genes: *Gata1, Cebpa, Runx1, Myb, Zfpm1*, and *Meis1*; Functional Genes: *Mpo* and *Gypa*; Supplementary Table 1) within phenotypically mixed Kit+Lineage- HSPC (Extended Data Fig. 1f). *PU.1* nascent transcription was higher in cells in a *PU.1* high state, as anticipated (Extended Data Fig. 1g,h). Furthermore, cells in the *PU.1* high state had more 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*, consistent with expectation. We found minimal change in the expression of *Runx1*. These experiments demonstrate the utility of smFISH in studying transcription in primary HSPC.

Next, we compared the sensitivity of smFISH to scRNAseq by comparing mRNA detection of the 7 aforementioned TF genes as determined by smFISH and as reported by 5 scRNAseq datasets of comparable HSPC[9,22–25]. All scRNAseq datasets had a marked increase in the number of non-expressing cells for all genes tested compared to smFISH (Extended Data Fig. 2a). We then calculated the Gini coefficient for each gene, which quantifies population dispersion of a variable of interest[26]. As capturing the underlying population dispersion for an mRNA is essential for assigning transcriptional states, under-estimating 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). In 6/7 genes tested, the Gini coefficient was lower when determined by smFISH compared to scRNAseq (Extended Data Fig. 2b). The sole exception was *Zfpm1*, which had a similar calculated Gini index value in 1/5 scRNAseq studies. We then calculated Gini coefficients for a larger set of other transcriptional regulatory genes from scRNAseq (Supplementary Table 2). Consistent with the findings of our initial test set of TF genes, the majority of genes in this list had Gini indexes >0.8 (Extended Data Fig. 2c). Furthermore, we found that these sensitivity restrictions significantly effect 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, Supplemental Figure 2a, Supplemental Table 3, and Supplementary Discussion 1). As such, scRNAseq may be fundamentally incapable of providing quantitative estimates of transcriptional noise during hematopoiesis.

PU.1 and *Gata* TF co-expression in HSPC

Given these results, we then used smFISH to evaluate the role of stochasticity in a central transcriptional network between *PU.1* and the GATA TFs *Gata1* and *Gata2*. *PU.1* and *Gata1* are critical to differentiation along the granulocute-monocyte (GM) and erythrocyte (Ery) lineages, respectively, and the direct interaction between these TF through an antagonistic toggle switch was the original model for GM/Ery fate decisions[16,17]. *Gata2* is abundant in early HSPC and may function similar to *Gata1* in these cells by antagonizing *PU.1* function, albeit at lower potency[27]. Additionally, *Gata2* primes HSPC to upregulate *Gata1* during terminal erythropoiesis, after which Gata1 is thought to shut off *Gata2* in a phenomenon described as the “Gata switch”[28,29]. Nevertheless, recent scRNAseq studies have either outright failed to detect progenitors co-expressing *PU.1* and *Gata1*[23], or only detected

co-expression in 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 (GMP), Megakaryocyte/Erythrocyte Progenitors (MEP), and Common Myeloid Progenitors (CMP), and assessed the expression of these TF by smFISH (Fig. 1d, Supplemental Figure 1). *Gata1* was high in MEP and low in GMP, while *PU.1* was high in GMP and low in MEP; *Gata2* was highest in CMP (Fig. 1 d,e, Extended Data Fig. 3). Interestingly, we noted that in all instances except *PU.1* in GMP and *Gata1* in MEP, mRNA count distributions were positively skewed with the majority of the probability mass below 50 copies of mRNA/cell. Frequency distributions of this type are typical of mRNAs produced in infrequent bursts of transcription[30]. 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. Strikingly, the majority of CMP expressed *PU.1* (97%), *Gata2* (96%), and *Gata1* (64%) (Extended Data Fig. 3c). Most importantly, greater than 60% of CMP had at least 1 mRNA for all three genes, as well as most MEP (45%) and GMP (89%) (Fig 1g, see Extended Data Fig. 4 and Supplementary Discussion 2 for discussion of false positives in smFISH).

We next asked if CMP 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 *Gata1/2* were mutually exclusive, the empiric frequency of CMP with simultaneous transcription sites, $f_{PU.1+Gata1+}$, should be lower than the frequency predicted by statistically independent firing, $f_{PU.1+} * f_{Gata1+}$. To 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 ~1.5–2 fold higher than $f_{Gata2+} * f_{Gata1+}$, consistent with the proposed model of *Gata* gene co-expression during erythropoiesis. These findings indicate that mutually exclusive transcription of the antagonistic TF *PU.1* and *Gata1/2* does not occur in CMP.

Stochastic transitions to transcriptional termini

Given these findings, we then performed stochastic simulations using the transcriptional parameters inferred from our CMP data to model the transcriptional behavior of each gene over time[21,31]. To refine our parameter fitting, we assigned CMP to 4 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 CMP; see Extended Data Fig. 5a–b and Methods). Ordering these states with diffusion pseudo-time estimation (Fig. 2a)[32] identified two branches emanating from the LES CMP cluster. Consistent with our prior analysis, while each branch in the pseudotime plot had differential transcriptional activity, active transcription sites for the "opposing" TF 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 and Supplementary Methods 4). Single-cell trajectories simulated using parameters for a given state closely approximated the transcriptional behavior 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 a CMP's lifespan *in vitro*[33]. The majority of trajectories transcribed hundreds of copies of *Gata2* over this time period irrespective of the transcriptional state (Fig. 2e). Simulated LES and G2H CMP also transcribed between 20 to 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 >99% of cells were "triple positive" (TP) at some point during the 12 hours simulation window (Fig. 2f). Furthermore, the majority of trajectories were TP for over half of the simulation time-frame (Fig. 2g).

We then asked if bifurcation into P1H and G1/2H states could occur stochastically from the LES state. Indeed, although each state's parameter set 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 where they would then adopt new transcriptional parameters. 9% and 18% of trajectories initialized in the LES state ended up in either the G1/2H or P1H terminus after one CMP lifetime, respectively, while 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 while transcriptional noise drives co-expression of antagonistic TF, stochastic and reversible transitions of noisy states can still efficiently bifurcate into *PU.1* and *Gata* high expressing states.

Mapping HSC state correlations through pedigree analysis

While the above results suggest significant transcriptional stochasticity in CMP, a critical question is whether such phenomena occur in HSC. Moreover, the effect such processes have on the transcriptional state dynamics of the *PU.1-Gata1* network in HSC is currently debated[33–35], and "transcriptional priming" has been suggested as putatively limiting the transcriptional states an HSC and its descendants can occupy[9,36,37].

We first asked if HSC co-expressed *PU.1* and the *Gata* genes. HSC progeny had robust expression of all three genes at a similar level to CMP, with >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 employed kin correlation analysis, an experimental approach pioneered in seminal work on mESC that utilizes the information embedded in pedigrees to infer the dynamics of transcriptional state transitions[38]. To that end, we followed HSC for 96 hours *ex vivo*, constructed pedigrees from each HSC, and used smFISH to assign transcriptional states to cells (Fig 3a, Extended Data Fig. 7, Supplemental Figure 2b; see Methods and Supplementary Methods 5 for details on state assignments). In addition to the 4 subpopulations identified in CMP (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 (Fig. 3b). We excluded these cell populations to focus on more immature HSPC.

First, we determined if individual HSC could generate progeny in multiple states. 27/117 of colonies contained only 1 predominant state type: 5/117 were G1/2H-dominant, 2/117 P1H-dominant, and 21/117 LES-restricted (Extended Data Fig. 8a). The frequency of colonies with any combination of 2, 3, or 4 or more states was ~45%, ~25%, and ~3%, respectively. 25% of mixed colonies 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 conditional on the presence of a cell in each state (Fig. 3c). While no two states were mutually exclusive in this analysis, the frequency of finding a colony with both G1/2H and P1H states was low (3/117). HSC colonies that produced any G1/2H progeny had a 10-fold reduction in the frequency of P1H cells, a 10-fold 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 2-fold. On the other hand, clones producing P1H cells had a 3-fold reduction in G1/2H cells and a 4-fold reduction of G2H cells.

Stochastic and reversible HSC transcription state dynamics

One scenario that could account for such behavior would be an irreversible switch in the transcriptional kinetics arising early in the pedigree. In such a case, one would expect cells at close generational distances (e.g. sister cells) 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 (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 employed KCA. KCA uses the correlation in endpoint transcription states and lineage distance between cells in a pedigree to infer the transition rates of those states (Figure 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 to 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 end-point of the experiment. Additionally, we noted that some transitions had little to no probability per generation, e.g. direct transitions from P1H to G1/2H or vice versa had ~0% probability. Moreover, entering the G2H state appeared to be prerequisite to entering G1/2H. The inferred transition probabilities were robust to a range of mRNA cutoffs between different states, implying 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, indicating that such phenomena influenced the inferred transition probabilities (Extended Data Fig. 10).

We next used these transition probabilities to model a spectrum of state transition behaviors ranging 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, and Methods). At all generational distances tested, both the Markov chain (Model II) and fully connected model (Model IV) had approximately 100% and 30% reductions in the predicted 3-cell state frequency error when compared to the irreversible (Model I) and partially irreversible (Model III) models, respectively (Fig. 4e(iii)). Moreover, the Markov chain performed better at lower generational distances (e.g. v=2). Overall, among the models tested, state transition models containing 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 wanted 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 behaviors in clonal offspring of single HSC 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 these three genes, may not be fully predictive of the past or future states visited by that cell's ancestors or offspring, respectively, even while 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. While "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 necessitating complex regulatory measures to facilitate such behavior, all of which will be similarly subject to molecular noise[7] (Supplementary Discussion 3).

In this paper, we have attempted to quantitatively address the question of TF gene expression noise in primary HSPC by utilizing single-molecule imaging and the quantitative analysis of pedigrees. Our results indicate that antagonistic TF are indeed co-expressed in the majority of HSPC, and that stochastic transitions between the transcriptional states defined by these genes are the likely basis of the system's dynamics (Figure 4i).

Methods

All reagents used in these studies are listed with catalog number in Supplementary Table 6.

Animal Husbandry

6–10 week old male and female C57/Bl6 mice 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 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. 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% Pen/Strep, 1% sodium bicarbonate, 74.8uM monothioglycerol and recombinant mouse (rm) SCF (50 ng/mL). The HPC-7 cell line was originally provided by Dr. Omar 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). KL populations (CMP, MEP, and GMP) were grown on retronectin coated (40ug/ml) #1.0 glass, 35mm^2 MatTek dishes in IMDM for with 1%Pen/Strep, 10% FBS and supplemented with recombinant mouse (rm) SCF (100 ng ml-1), rmTPO (100 ng ml-1), rmIL-3 (10ng ml-1), rmIL-6 (10ng ml-1), and recombinant human (rh) EPO (2IU/mL) and GM-CSF (10ng/mL). M-CSF (10ng/mL) and G-CSF (10ng/mL) was supplemented to GMP cultures. Bulk KL cells used in Figure 1 and Figure 2 were grown in suspension in a single well of a 24 well plate.

Cells were grown for ~12–16 hours *ex vivo* to allow for full recovery from sorting prior to analysis with smFISH. HSC in Fig S5 were grown for 72 hours on retronectin coated MatTek dishes, as above, in StemSpan SFEM media with 1% Pen/Strep and recombinant mouse (rm) SCF (100 ng ml-1), mTPO (100 ng ml-1), rmIL-3 (10ng ml-1), rmIL-6 (10ng ml-1), and rhEPO (2IU/mL). Cells were maintained at 37°C and 5% CO_2.

Onstage Culture

For time lapse imaging, sorted HSC were seeded on 35mm^2 MatTek dishes coated with 10ug/ml anti-CD43 biotin instead of retronectin in order to reduce cell movement and cell loss/misidentification during the experiment[39]. 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

5–10 mice 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, 1mM EDTA) and

filtered through a 70μm filter. Red blood cells and granulocytes were then removed through density centrifugation over a 5mL Histo-Paque Ficoll Gradient. After extensive washing of the buffy coat, cells were then lineage depleted using 1:1000 dilution of anti-mouse B220, CD19, CD4, CD8, Gr-1, CD11b, Ter119, and CD127, all biotinylated, on ice for 25 minutes. Cells were washed and then stained with triple washed anti-IgG magnetic beads (Untouched Mouse T Cells Kit, ThermoFisher) on ice for 30 minutes. 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 minutes on ice with anti-CD150, anti-CD34, anti-CKIT, anti-Sca1 and anti-CD48 (all 1:250) and anti-CD16/32 (1:500) with Streptavidin Pacific Orange (1:1000). Cell populations were sorted on 4-way purity mode into IMDM, 5% FBS, 1%Penn/Strep. See Supplemental Methods for gating strategy.

Poly-L-Lysine Coating of #1.0 12mm Coverslips

To prepare poly-L-lysine coated coverslips for immobilization of suspension cells (HPC-7, Kit+Lineage- Progenitors, and whole bone marrow), 12mm #1 Coverslips were first boiled in 0.5N HCl for 30 minutes, washed extensively in double distilled water and stored in 70% ethanol. Coverslips were then coated for 5 minutes with 0.01% poly-(L)-lysine, followed by two washes with water and air-dried for 20 minutes. Coverslips were then transferred to a 24 well dish on ice for cell immobilization and subsequent smFISH staining. 20uL cell aliquots of ~10,000 cells/100uL were dotted and spread onto the coverslip and the cells were allowed to settle on ice for 20 minutes. Unstuck cells were then washed away with 2 washes in PBS prior to 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 of duplexes >−0.1kJ/mol; ΔG of Hairprin formation >−0.1kJ/mol.

Putative sequences were then screened for off target activity using Blastn (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” length of 70mer. Probes were then ordered in 100nmol quantities from Thermo Fisher or IDT. Individual probes were resuspended at 100uM concentration, mixed in equal proportions to 10uM final concentration of each probe, and stored at −20°C. Stock solutions were diluted in Ultrapure water to 200ng/uL 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 10nM final concentration.

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

Cells were fixed in 3.2% PFA (Electron Microscopy Sciences) diluted in PBS with 1mM MgCl2 (PBSM) at RT for ten minutes. Cells were then washed with 2 ml cold PBSM with 10mM Glycine. Cells were then permeabilized on ice for 20 minutes in PBSM with 0.1% Triton-X1000 and 2mM Vanadyl Ribonucleoside Complex (VRC). After washing with PBSM, cells were then incubated at RT 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, 2mM VRC, 10ug/ml sheared ssDNA from salmon sperm, 10ug/ml *E.coli* tRNA, 10ug/ml molecular grade bovine serum albumin, and 200ng each of 70mer primary probe mixes. Cells were then washed twice for 20 minutes at 37°C with prehyb-30, and once with 2X SSC. Cells were then post fixed in 1% PFA in PBSM for 5 minutes, followed by two washes in 2X SSC. Primary stained cells were then washed with prehyb-10 (10% formamide, 2X SSC) for 10 minutes at 37°C and stained with 10% Dextran Sulfate, 10% formamide, 2X SSC, 2mM VRC, 10ug/ml sheared ssDNA from salmon sperm, 10ug/ml *E.coli* tRNA, 10ug/ml molecular grade bovine serum albumin, and 10ng each of 20mer readout probes for each gene for 3 hours at 37°C. Reactions were then washed with 2 washes of 10 minutes 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 12mm #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 Supplemental Table 5 for probe sequences.

Probe Labeling with Cy3, Cy5, and Alexa Fluor 594

Secondary “readout” probes were purchased from Thermo Fisher with 5’ C5 amine and 3’ C7 amine modifications. 5 ug of each readout probe was then coupled to the appropriate fluorescent dye according to the manufacturer’s specifications. After labeling, probes were extracted from excess dye by use of a Qiagen Nucleotide Removal Kit, resuspended in Ultrapure water, and stored at −20°C. Labeling efficiency was determined using Beer’s law. Only fluorescently labeled probes with >1.5 dyes/oligonucleotide were utilized 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 600ms, 600ms, 400ms, and 10ms, 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, Inc.).

Time lapse Microscopy

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

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 @3.20GHz and 32Gb RAM.

Image Analysis: smFISH

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

Probabilistic Transcriptional State Assignments

See Supplemental Methods.

Comparison of scRNAseq and smFISH

See Supplemental Methods.

Summary Statistics of mRNA copy number/cell

Extended Data Fig. 3c provides summary statistics for the mRNA counts/cell for *PU.1, Gata1*, and *Gata2* in primary KL populations. N indicates the total number of cells analyzed across two separate experiments (CMP and MEP) or in a single experiment (GMP). μ is the arithmetic mean of the number of mRNA molecules/cell, 95% CI is the 95^{th} confidence interval and %Expressing are the number of cells with >=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 Neighbor Embedding Maps (tSNE)

tSNE maps of primary KL cells were generated in MATLAB with the ‘tsne’ function using the mature and nascent mRNA values/cell for each gene as variables.

Transcriptional State Assignments: KL

Gating strategy is shown in Extended Data Fig. 5. From all CMP, large, polyploid megakaryoblasts with hundreds of copies of all three genes are first remove. Next, all cells with *Gata1* >10 are classified as G1/2H (red box, top left histogram). The negative fraction (gray 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*<50. A small population of $Gata2^{high}/PU.1^{high}$ CMP (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} and k_{ini} for *PU.1*, and an increase in the k_{off} and k_{ini} for *Gata1*. 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/CMP cell for *PU.1, Gata1*, and *Gata2* were generated in MATLAB using the Diffusion Pseudotime Estimation software described by[32]. The DPT maps were generated using a 40-nearest neighbor search with a kernel width of 50. The diffusion map plotted in Figure 2a are the first two diffusion components and is colored according to the transcription state classification scheme described in Extended Data Fig. 5. For the raster spike density plots in Figure 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 *PU1 Branch*, we subsetted on the cells in the LES and P1H states. Each spike is a cell and the height of the spike is the number of active TS in that cell.

Phase Portrait Diagrams

Phase portraits are based on similar analyses by[40], with the nascent mRNA/cell for a gene given on the y-axis and the mature mRNA for a gene given on the x-axis. >240 cells were analyzed 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 TS.

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 labor intensive. Imaging the entire surface of the dish for smFISH requires ~500 stage positions with a 100X objective, which is prohibitively long for four color 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 where we could confidently identify colonies on the epifluorescent microscope and imaged these colonies for smFISH. We then manually analyzed the time lapse movies for these select colonies in TTT[41]. 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 *Gata1*>10 copies/cell. P1H macrophage cells are all cells with *PU.1*> 150 copies/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 panel). As such, we fit both genes to a two-component negative binomial (NB) distribution. For the data in Figure 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* >80%, respectively. For the transition dynamics shown in Figure 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 the original KCA paper[38]. 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 NB components.

KCA—An elegant and rigorous derivation of KCA can be found in[38]. Scripts to perform KCA and consistency checks were adapted from scripts generously provided by Dr. Sanand Hormoz and Dr. Michael Elowitz and are provided in the online supplement along with the raw data for all colonies analyzed.

Briefly, KCA was performed using all colonies analyzed 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 Figure 4b–c are average inferred transition probabilities for each lineage distance u, and the error bars are the standard error in those estimates derived by bootstrapping through the data 5000 times. The script entitled “KCA.m” will generate all the figures found here plus will save the mean and std of the inferred transition probabilities between states.

Checking Robustness of mRNA Cutoff Threshold—We employed the approach formulated by Hormoz and colleagues[38], 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 employed in the KCA to analyze the correlation in mRNAs between cells recently having divided within the last hour prior to fixation at the endpoint. That analysis also revealed considerably high correlation in mRNA values.

Taken together, these results indicate that our results are likely not significantly affected by partitioning asymmetries of mRNAs during mitosis.

Comparing Reversible and Irreversible Dynamics—To test whether our data was better described by dynamical models containing irreversible transitions (Model I and III) versus those without (Model II and IV) we used an elegant approach described by Hormoz et al. 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(ii) 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 1000 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 modeling 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 employed 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 Material).

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 Figure 1b, the gramm package. Exported .EMF or .JPEG 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.

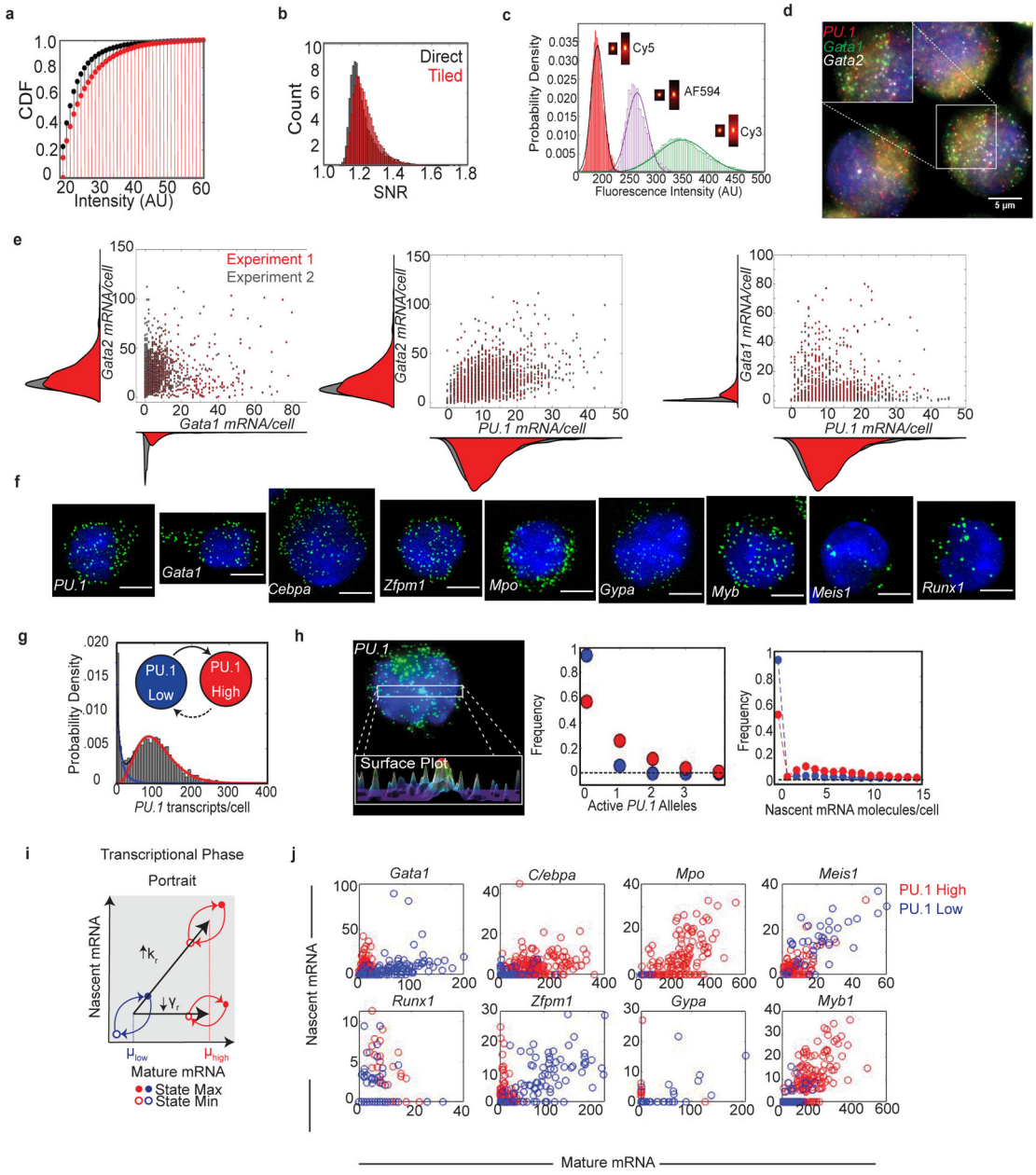
Data availability

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 Fig. 8 and 9), separate .mat files have been provided as Supplementary Data 1 and has also uploaded to the Github repository listed above or are generated upon running the associated scripts.

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 Figures 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. Finally, all software generated by other groups used in this study are listed in Supplementary Table 7.

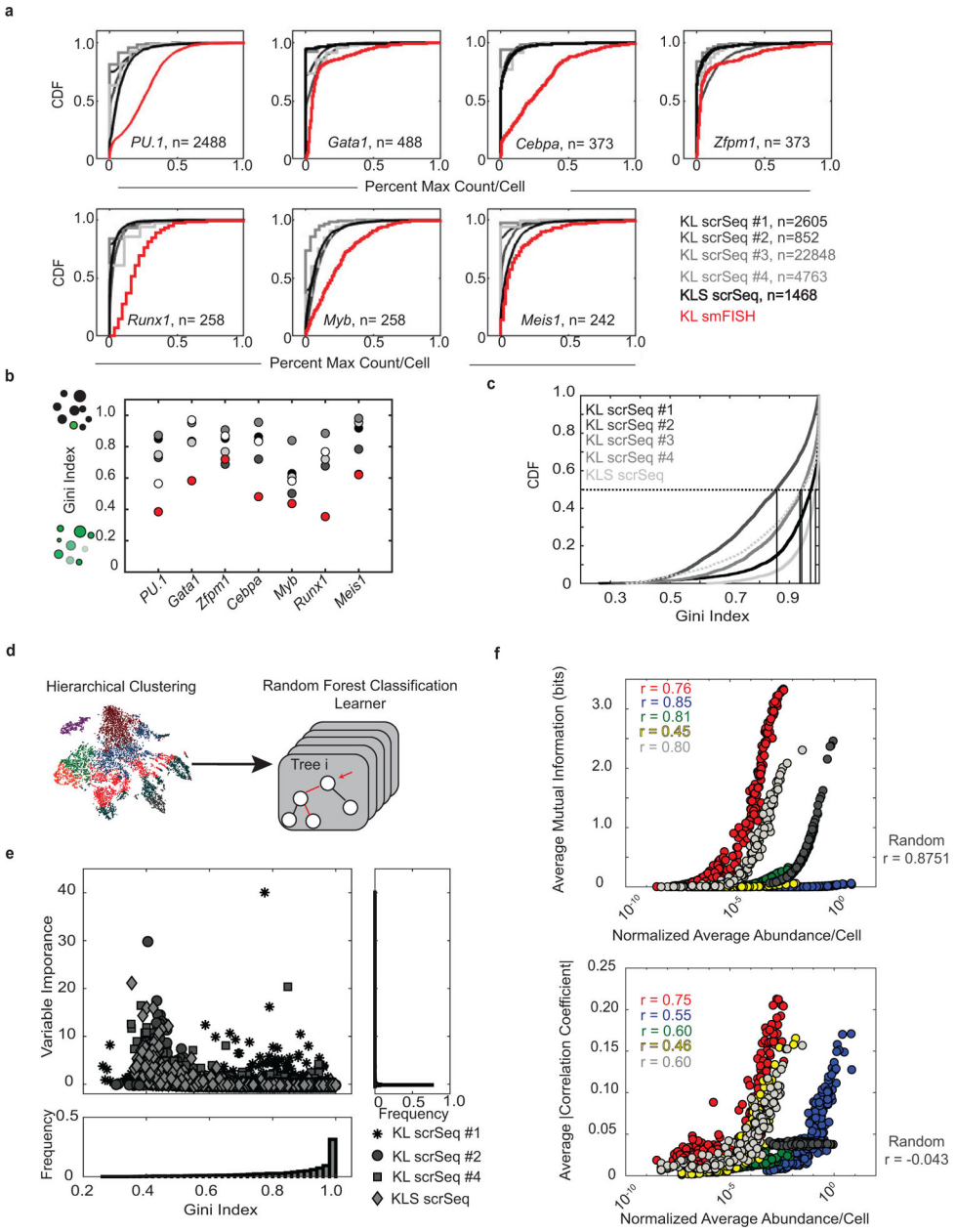
Extended Data



Extended Data Fig. 1|. Transcriptional dynamics of genes conditional on *PU.1* state.

a-b, Among all spots that passed intensity and 3D-PSF fit thresholding in FISH-QUANT, **(a)** Cumulative Distribution Function (CDF) of spot intensity and **(b)** histogram of signal to noise in spot intensity to local background intensity. **c**, Probability densities for fluorescence/mRNA molecule in HPC-7 cells for Cy3, AlexaFluor 594, and Cy5 labeled readout probes. Insets are XY and XZ average PSFs for each fluorophore. Overlaid line is fit to Gaussian distribution. >10,000 spots per fluorophore. **d**, Representative three color smFISH for *PU.1* (Cy5, red), *Gata2* (Cy3, white) and *Gata1* (AF594, green) in HPC-7 cells. Scale bar = 5um. **e**, Bivariate distributions of *Gata1-Gata2* (left), *Gata2-PU.1*

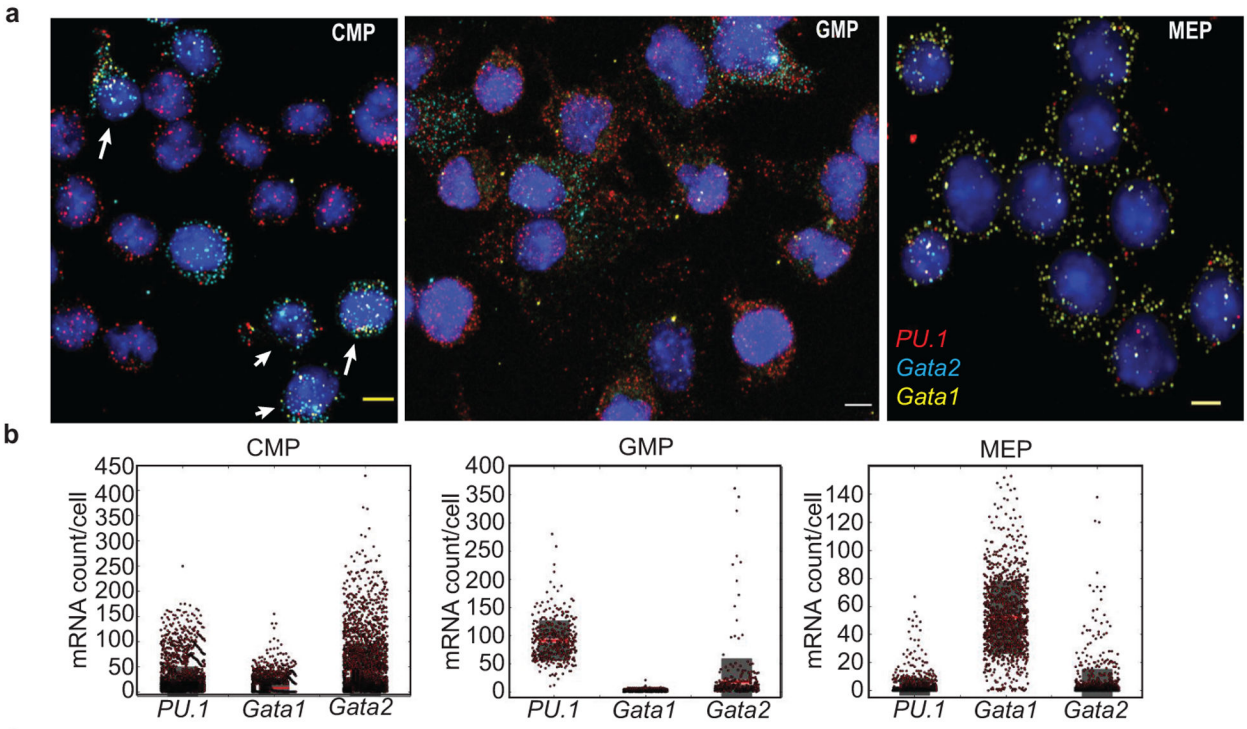
(middle), and *PU.1-Gata1* (right) in two independent experiments (n>400 cells/experiment) with HPC7 cells. **f**, Representative images of multiplexed smFISH between *PU.1* and 8 other hematopoietic genes in Kit+Lineage- bone marrow from wildtype mice (n=258–2488 cells for each gene; derived from single experiment; scale bar 5um). **g**, Probability distribution for *PU.1* mRNA/cell in KL cells from wildtype BM. 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 imaging of PU.1 smFISH with a single, large transcription site in the nucleus. (Middle) frequency of cells with indicated number of active *PU.1* transcription sites. (Right) Frequency distribution of summed nascent mRNA/cell in each PU.1 state. **i**, Schematic demonstrating a hypothetical transcriptional phase portrait. **j**, Phase portraits for each gene based on the cell's *PU.1* state.



Extended Data Fig. 2|. Comparative Analysis of smFISH and scRNAseq.

a, CDF plots of mRNA/cell for 5 scRNAseq datasets and smFISH. Data is normalized to the max count for each gene in each data set. **b**, Calculated Gini index for 7 TF mRNAs in each scRNAseq data set (white through black) and smFISH (red). **c**, CDF plots of Gini index for all 5 scRNAseq datasets (See Supplemental Table 2 for gene list). **d**, Schematic of Hierarchical Clustering followed by Random Forest classifier to identify important variables for cluster assignment. **e**, Gini coefficient versus variable importance for 4 scRNAseq datasets. Bottom and right panels are marginal distributions of Gini and VI, respectively. **f**, Plot of average mutual information (MI, top) or average absolute value of the Pearson's correlation coefficient (PCC, bottom) versus normalized abundance of n=200 randomly selected genes against all other genes in the dataset. R values listed are the correlation

coefficients between abundance and MI or PCC. See Supplemental Discussion for further details on the analyses performed.



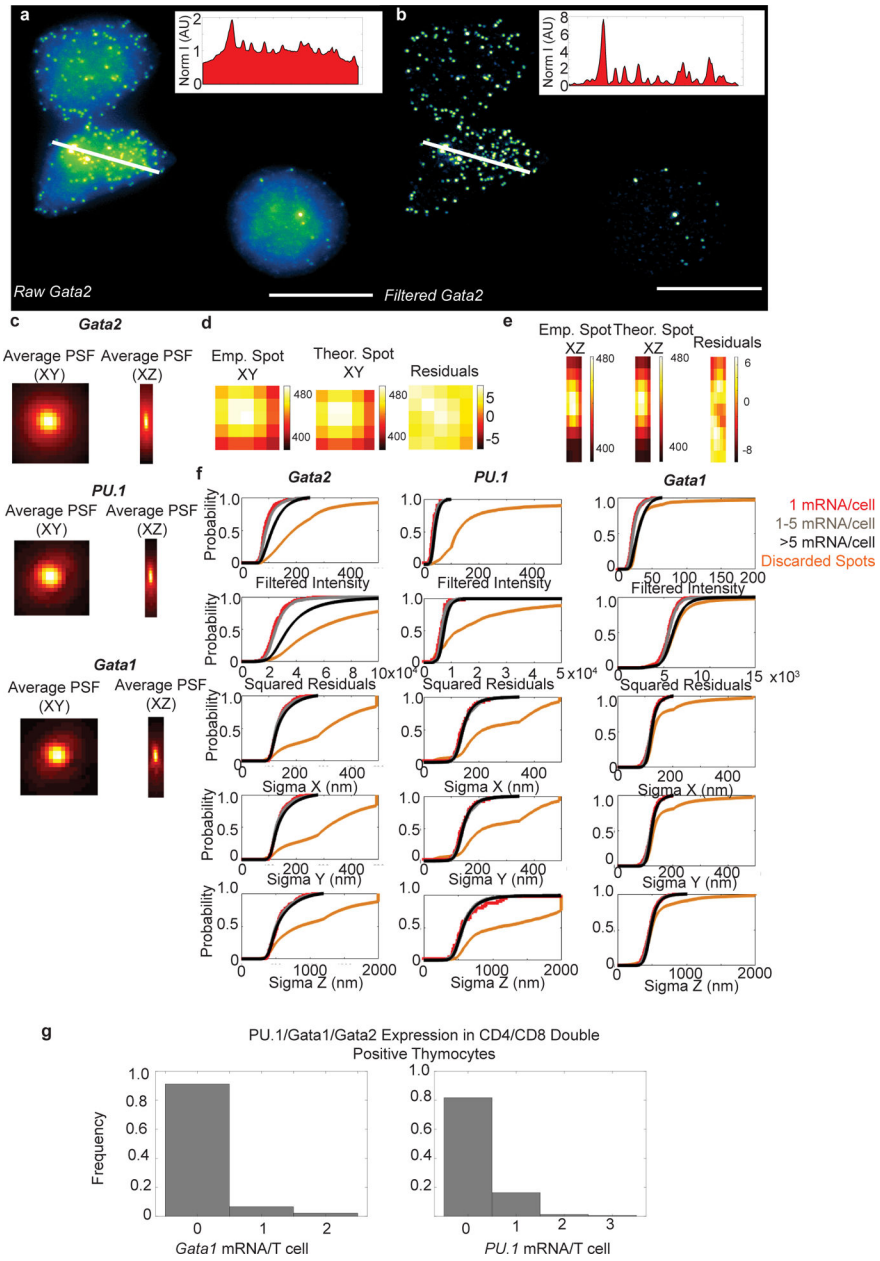
c

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%

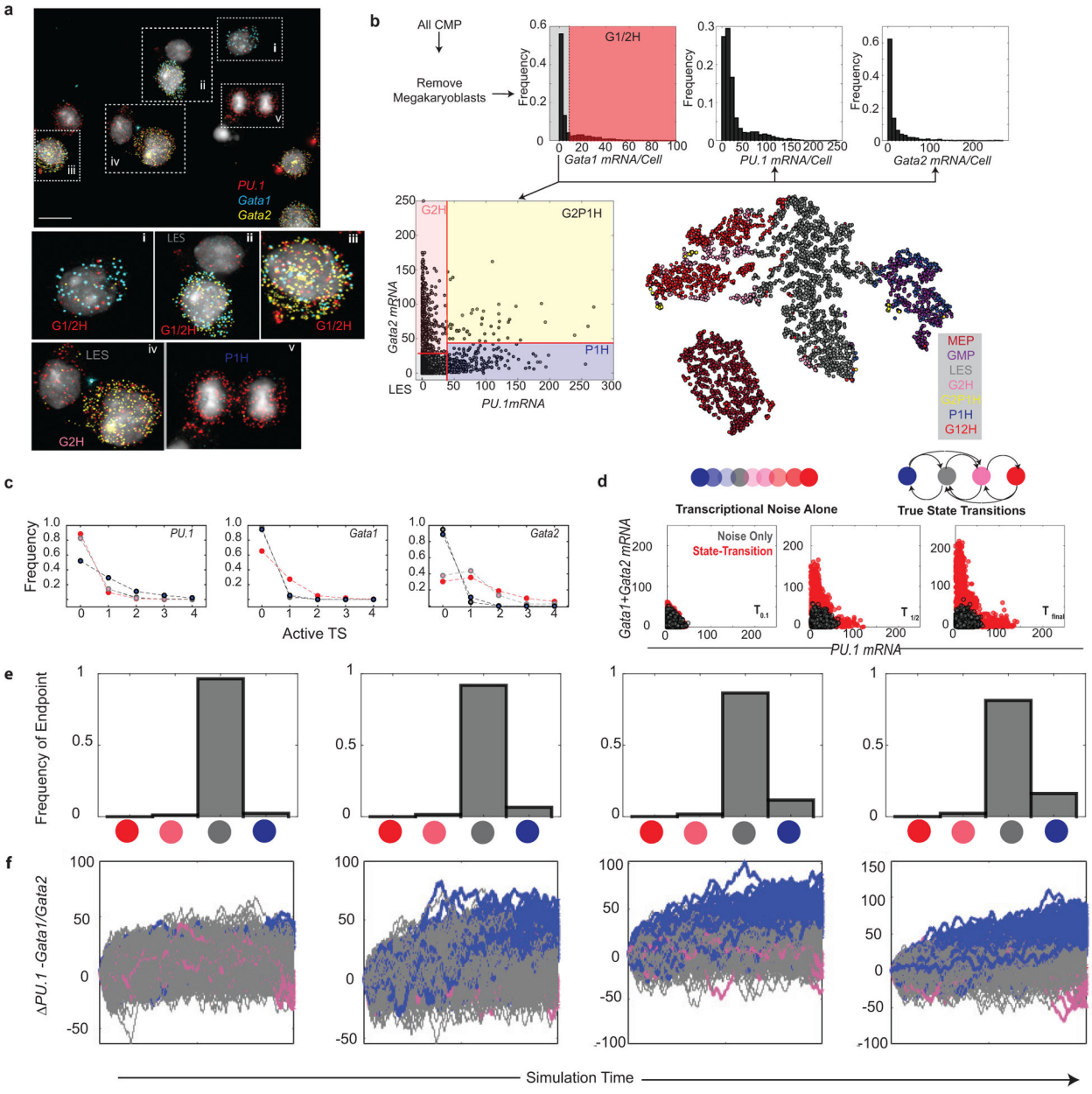
Extended Data Fig. 3|. Summary statistics of mRNA copy number for primary KL.

a, Representative images of CMP, GMP, and MEP cells stained by smFISH for *PU.1/Gata1/Gata2*. Scale bars= 5μm. Arrows point to CMP co-expressing all three mRNAs. **b**, Boxplots for mRNA count/cell with overlaid single cell mRNA values (dots). Gray box is 95 percent confidence interval, red line is mean expression, pink box is +/−SEM. **c**, Table of summary statistics for each gene. Data for **(a-c)** derived from two experiments (CMP and MEP) or a single experiment (GMP). Sample size is listed in the table 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 image from CMP (representative of >2 experiments in CMP; spot quality consistent with all reported experiments in this manuscript). Insets are line intensity plots (indicated on cell in white). Scale bar is 10μm. **c**, Average point spread function (PSF) in XY (left columns) and XZ (right columns) for each gene from all detected spots from CMP dataset. **d-e**, Empiric (left) versus theoretical (middle) PSF and residuals (right) in the XY (d) and XZ (e) planes. **f**, Cumulative distribution functions for all spots passing the initial intensity thresholding for filtered intensity (top row), squared residuals (2nd row), and width of spots in X, Y, and Z in nanometers (3rd-5th row, respectively). Spots are separated based on those coming from cells with >5 copies of mRNA/cell, between 2–5 copies/cell, and 1 copy/cell. Discarded

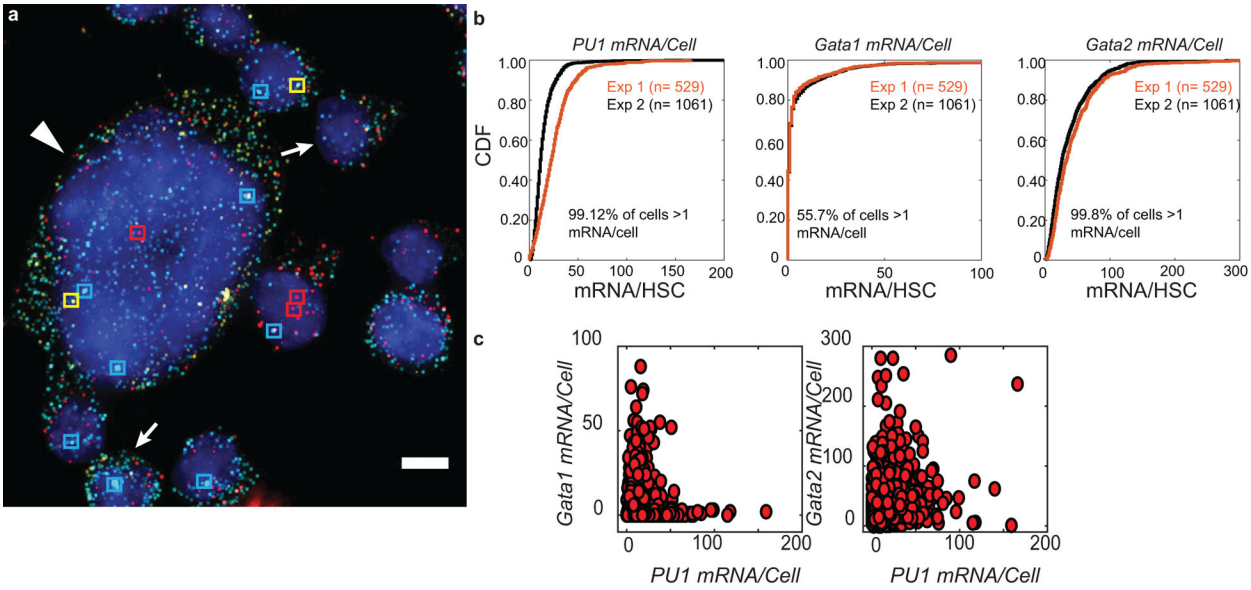
spots failing 3D fitting are shown in orange. **g**, mRNA detection in primary CD4+/CD8+ thymocytes (n= 136 for *Gata1*, n = 154 for *PU.1*).



Extended Data Fig. 5|. Gating strategy to assign CMP to states.

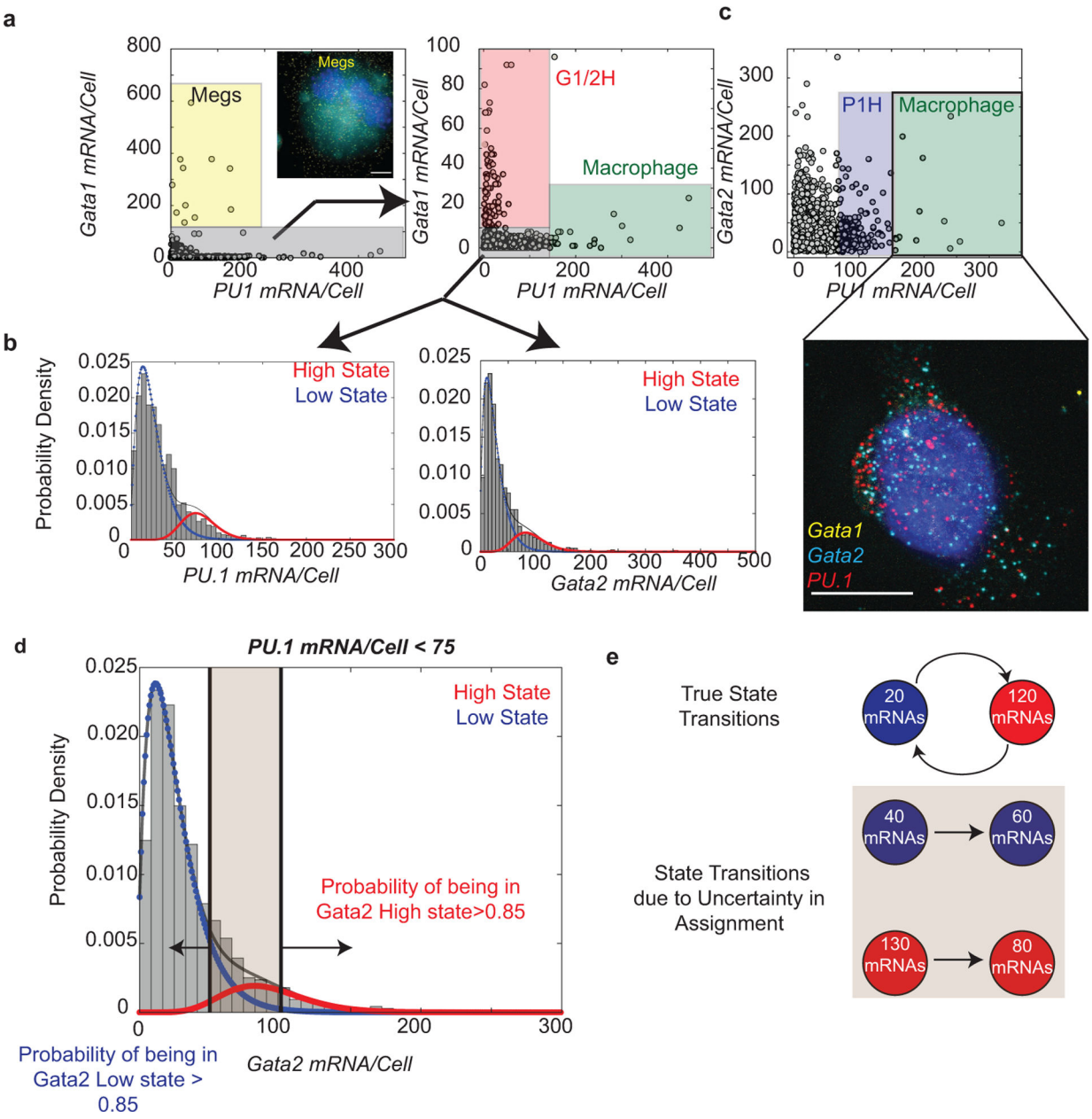
a, Representative images of CMP in different states. Scale bar = 10μm. **b**, Gating scheme for assigning CMP to transcriptional states. See Supplementary Discussion for details on the gating strategy. tSNE plot demonstrates the proximity of states to one another and to immunophenotypic GMP and MEP. Images and analyses derived from experimental datasets reported in Fig. 1 and Extended Data Fig. 2. c, Frequency distribution of transcriptional bursting for each gene in each transcriptional state. 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 (edges). (bottom) Time dependent behavior of simulated cells in a noise only (gray) or state transition system (red) shown as a bivariate plot of *PU.1* copy number versus *Gata1+Gata2* copy number. T indicates the amount of 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 only changes the mRNA half-life of the mRNA defining that state. **e**, Endpoint states reached in the simulations (n=10,000) and **f**, 1000 representative simulation trajectories, color coded on the final endpoint state. Each panel is a different factor change in the mRNA half-life, with the left-most panel as the reference (i.e. the half-lives used in Fig. 2), 2X (second panel from left), 3X (second from right), and 4X (right-most).



Extended Data Fig. 6|. 72-hour progeny of HSC.

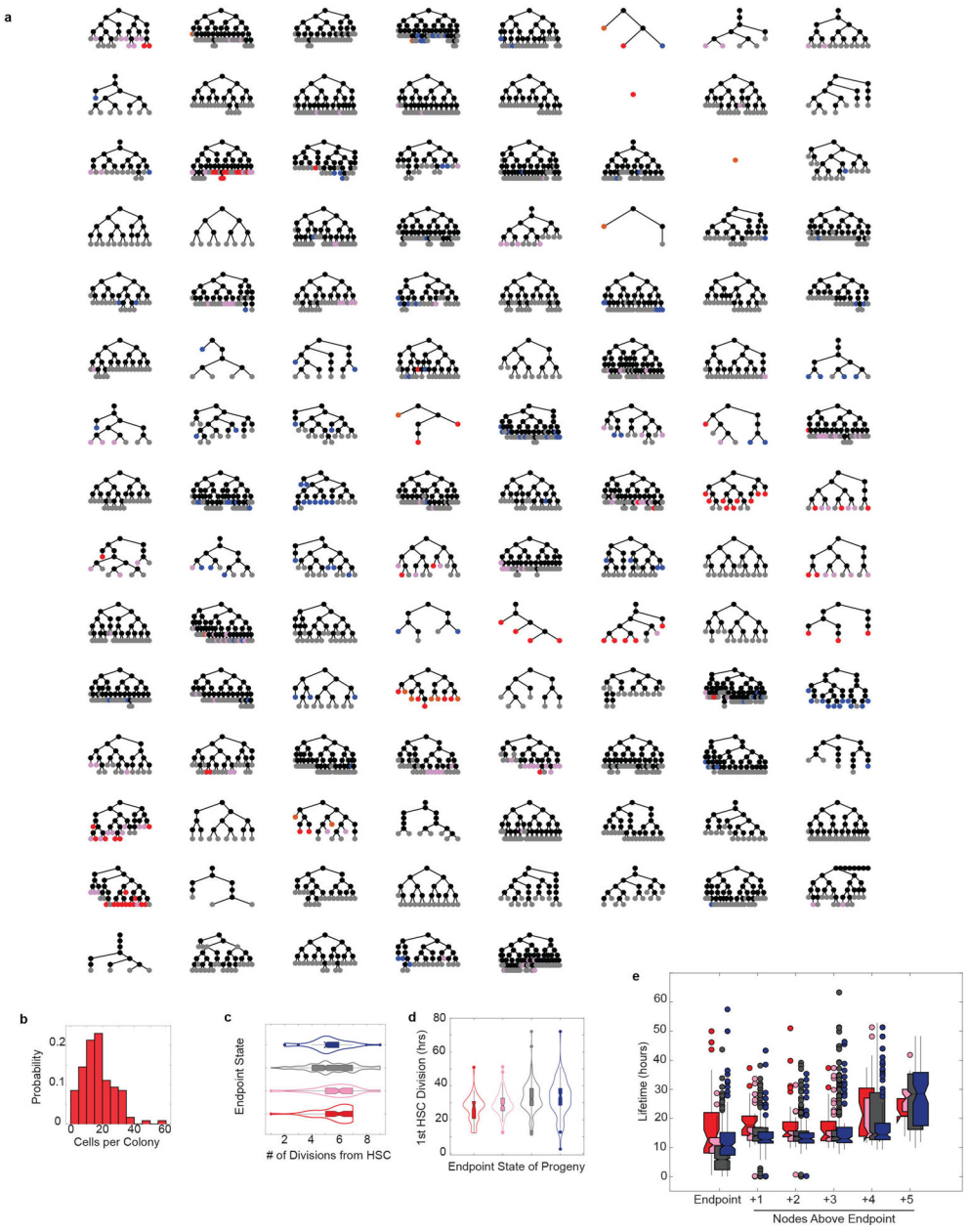
a, Representative images of HSC progeny. *PU.1* in red, *Gata2* in cyan, *Gata1* in yellow. Transcription sites are demarcated with boxes. Arrows are triple positive cells. Arrow head is a megakaryocyte. Representative of two separate experiments. **b**, CDFs for mRNA counts/HSC progeny. Number of cells with > 1 mRNA/cell is indicated. 2 separate experiments, (Exp1, n = 529; Exp 2, n = 1061). **c**, Bivariate distributions of *PU.1* versus *Gata1* and *PU.1* versus *Gata2*.



Extended Data Fig. 7|. State Assignments for HSC progeny.

a, Gating strategy. (left) Removal of megakaryocytes occurs first. (middle), Cells with >10 copies of *Gata1* are assigned to G1/2H. while cells with >200 copies of *PU.1* are assigned to P1H. **b**, Probability density distributions for *PU.1* and *Gata2* with overlaid fits for a two-component negative binomial distribution amongst cells after removing Meg-, G1/2H, and P1H with *PU.1*>200copies. **c**, Bivariate distribution of same cells. Contrary to the case in CMP, 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. 2). As such, all cells with *PU.1*>75 were assigned to P1H. **d**, Probability distribution for *Gata2* in remaining cells, fit with a two-component negative binomial. Such a distribution cannot be definitively separated into

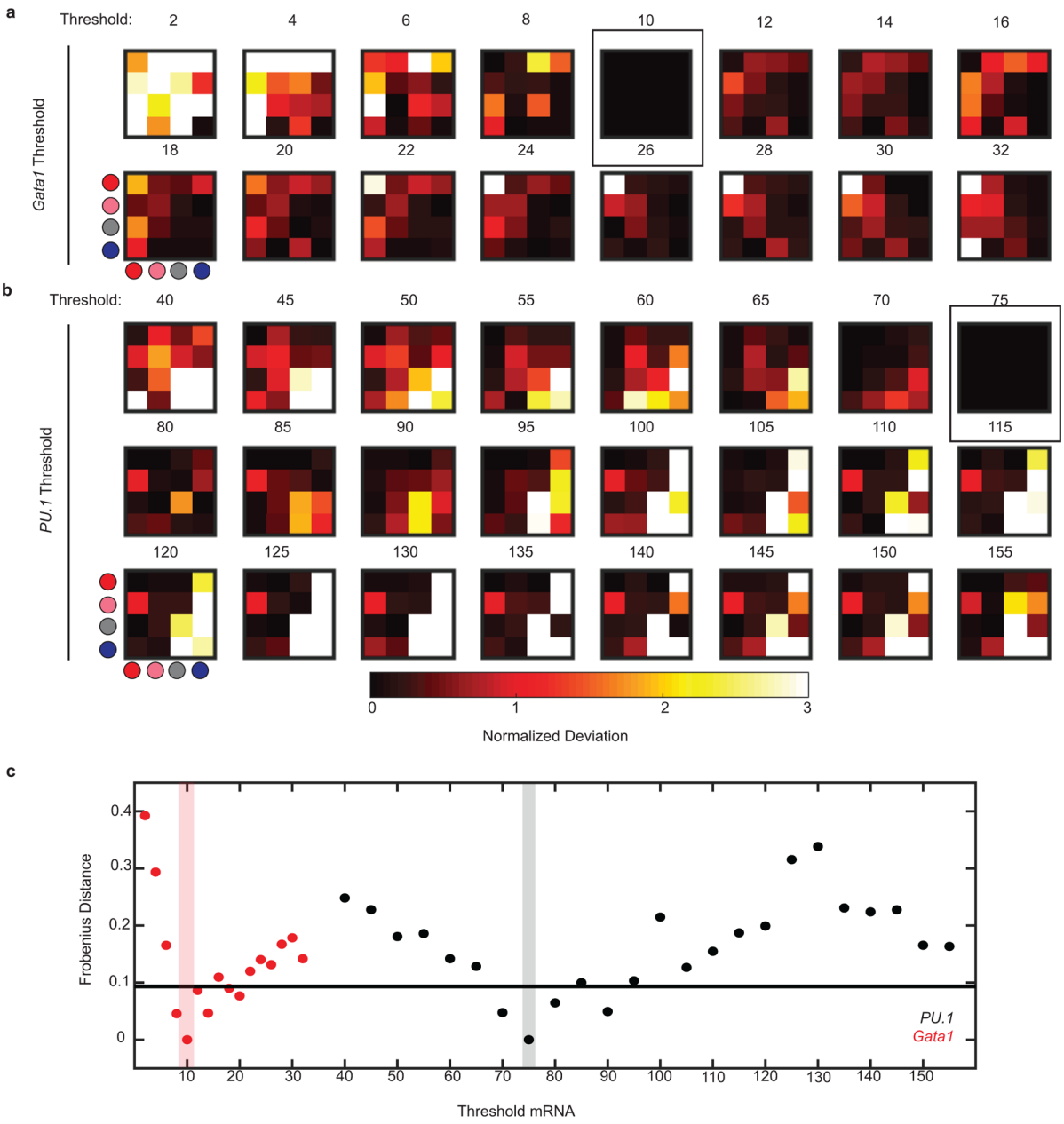
high and low components due to overlap in the distributions; therefore, cells are assigned probabilistically during KCA to the G2H or LES state in order to correct for false transitions arising from uncertainty in the assignment **(e)**. See Supplemental Discussion for more details on the rationale and implementation of probabilistic gating.



Extended Data Fig. 8|. HSC colony data.

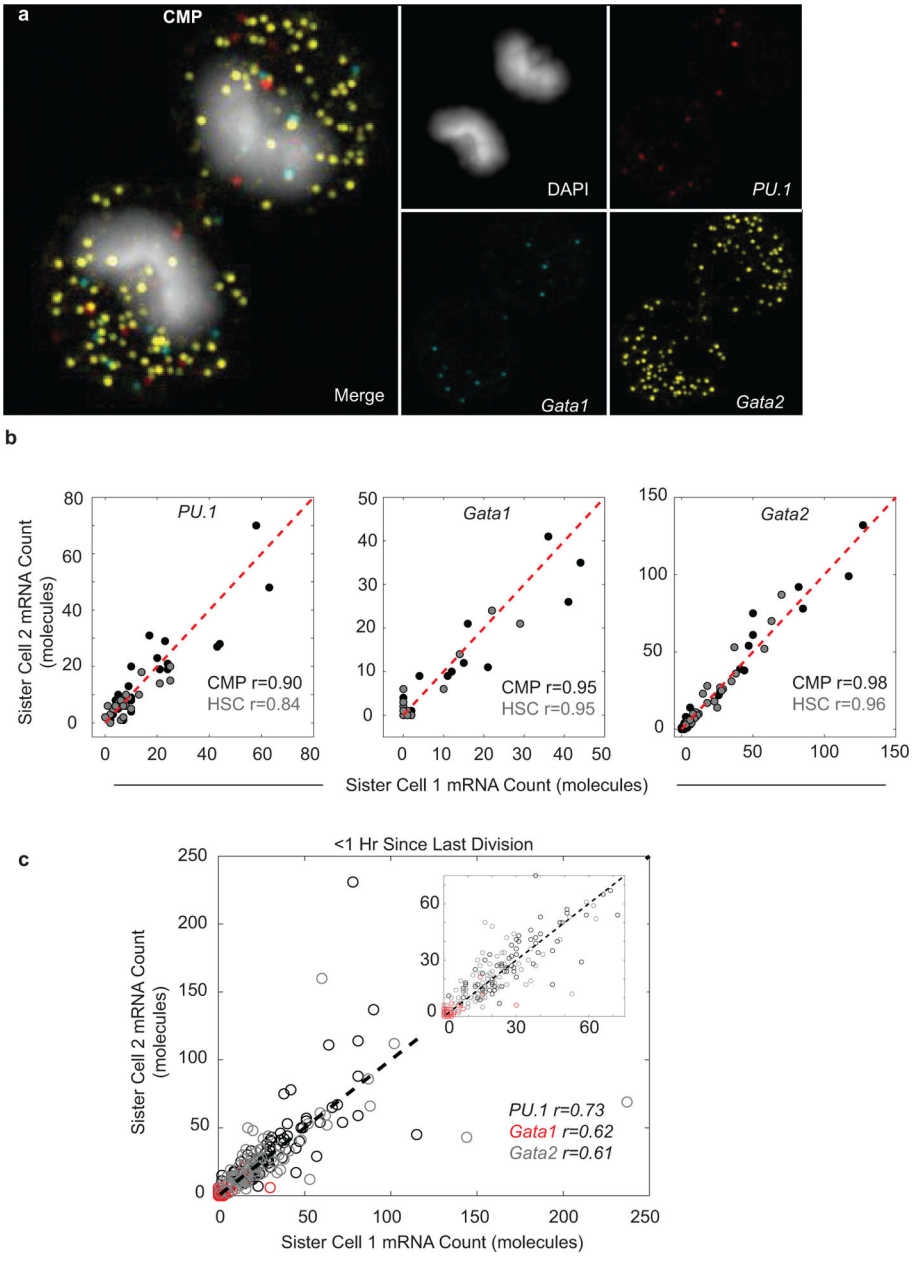
Endpoint cells are the leaves on each pedigree. Note that edge lengths are not scaled on time between divisions, and all endpoint cells are 96 hours from the start of the experiment. Cells are color coded consistent with the color scheme used throughout the manuscript. Megakaryocytes are labeled in orange. Nodes (cells) observed upstream of the endpoint (i.e. no transcriptional data is available) are colored black. **b**, Histogram of number of progeny

from a single HSC **c-e**, Proliferation phenotypes of cells based on end point state identity (P1H n=137; LES n=1571; G1/2H n = 81; G2H n =166). Cell lifetimes in **(e)** are time interval between cell birth (last division) and the next cell division or cell death. Violin plots are normalized to area with center box-and-whisker showing the mean, standard deviation and 95% confidence interval. Box-and-whiskers in **(e)** are mean, standard deviation and 95% confidence interval, with single dots representing outliers in 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/cell relative to the reference matrix reported in this manuscript (cutoff = 10 mRNA/cell). Boxed matrix is the reference

matrix. For any given transition (i.e. matrix entry), the initial states are the columns, final states are rows. Color code is same as used elsewhere in the manuscript. **b**, Same as in **(a)** except for *PU.1* (cutoff in manuscript = 75 mRNA/cell). **c**, Frobenius distance (FD, $\sqrt{\sum_{ij}\left(T_{i,j_{ref}} - T_{i,j_{test}}\right)^2}$) between each matrix versus the reference transition matrix. Solid black line indicates the background FD derived from statistical uncertainty in the reference transition matrix, derived by bootstrapping through the analysis n = 1000 times and picking random transition rates from a Gaussian distribution defined by inferred mean and standard deviation of the transition matrix. FD 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 CMP (n=52) and HSC (n=46). Pearson's correlation coefficient for sister cell mRNA copy number. Red dashed line is y=x. **c**, Correlation in mRNA levels between HSC that divided within the last 1 hour (n=171). Pearson's correlation coefficients for each gene are listed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments.

We thank Dr. David Shechter, Dr. Kira Gritsman, Dr. Robert Coleman, Jeetayu Biswas, Evelina Tutucci, Maria Vera Ugalde, and Richard Pisczcatowski for discussions; Dr. Florian Mueller for assistance with FISH-QUANT; Dr. Michael Elowitz and Dr. Sanand Hormoz for the scripts used for KCA; Melissa Lopez-Jones for assistance in probe design; Dr. Dirk Loeffler and Dr. Timm Schroeder for input on time lapse imaging of HSC; and Dr. Daqian 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 JCW, R01CA217092 to US, and U01DA047729 to RHS. US was supported as a Research Scholar of the Leukemia and Lymphoma Society and he 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.

References

1. Levsky JM & Singer RH Gene expression and the myth of the average cell. Trends Cell Biol. 13, 4–6 (2003). [PubMed: 12480334]
2. Elowitz MB, Levine AJ, Siggia ED & Swain PS Stochastic gene expression in a single cell. Science 297, 1183–1186 (2002). [PubMed: 12183631]
3. Raser JM & O'Shea EK Control of stochasticity in eukaryotic gene expression. Science 304, 1811–1814 (2004). [PubMed: 15166317]
4. Bar-Even A et al. Noise in protein expression scales with natural protein abundance. Nature Genetics 38, 636–643 (2006). [PubMed: 16715097]
5. Gandhi SJ, Zenklusen D, Lionnet T & Singer RH Transcription of functionally related constitutive genes is not coordinated. Nat Struct Mol Biol 18, 27–34 (2011). [PubMed: 21131977]
6. Huh D & Paulsson J Random partitioning of molecules at cell division. Proc Natl Acad Sci U S A 108, 15004–15009 (2011). [PubMed: 21873252]
7. Lestas I, Vinnicombe G & Paulsson J Fundamental limits on the suppression of molecular fluctuations. Nature 467, 174–178 (2010). [PubMed: 20829788]
8. Olsson A et al. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. Nature 537, 698–702 (2016). [PubMed: 27580035]
9. Tusi BK et al. Population snapshots predict early haematopoietic and erythroid hierarchies. Nature 555, 54–60 (2018). [PubMed: 29466336]
10. Femino AM, Fay FS, Fogarty K & Singer RH Visualization of single RNA transcripts in situ. Science 280, 585–590 (1998). [PubMed: 9554849]
11. Torre E et al. Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH. cels 6, 171–179.e5 (2018).
12. Chen KH, Boettiger AN, Moffitt JR, Wang S & Zhuang X RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 348, aaa6090 (2015). [PubMed: 25858977]

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). [PubMed: 27599845]
14. Chen HM, Pahl HL, Scheibe RJ, Zhang DE & Tenen DG 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). [PubMed: 8096519]
15. Koschmieder S, Rosenbauer F, Steidl U, Owens BM & Tenen DG Role of transcription factors C/EBPalpha and PU.1 in normal hematopoiesis and leukemia. Int. J. Hematol 81, 368–377 (2005). [PubMed: 16158816]
16. Rekhtman N, Radparvar F, Evans T & Skoultchi AI Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells. Genes Dev. 13, 1398–1411 (1999). [PubMed: 10364157]
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). [PubMed: 11023493]
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). [PubMed: 15146183]
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). [PubMed: 17041602]
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). [PubMed: 26343801]
21. Skinner SO et al. Single-cell analysis of transcription kinetics across the cell cycle. eLife 5, e12175 (2016). [PubMed: 26824388]
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). [PubMed: 29915358]
23. Paul F et al. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell 163, 1663–1677 (2015). [PubMed: 26627738]
24. Nestorowa S et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood 128, e20–31 (2016). [PubMed: 27365425]
25. Tabula Muris Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature 562, 367–372 (2018). [PubMed: 30283141]
26. Torre EA et al. A Comparison Between Single Cell RNA Sequencing And Single Molecule RNA FISH For Rare Cell Analysis. (2017) doi:10.1101/138289.
27. Chou ST et al. Graded repression of PU.1/Sfpi1 gene transcription by GATA factors regulates hematopoietic cell fate. Blood 114, 983–994 (2009). [PubMed: 19491391]
28. Doré LC, Chlon TM, Brown CD, White KP & Crispino JD Chromatin occupancy analysis reveals genome-wide GATA factor switching during hematopoiesis. Blood 119, 3724–3733 (2012). [PubMed: 22383799]
29. Grass JA et al. GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. Proc. Natl. Acad. Sci. U.S.A 100, 8811–8816 (2003). [PubMed: 12857954]
30. Singer ZS et al. Dynamic heterogeneity and DNA methylation in embryonic stem cells. Mol. Cell 55, 319–331 (2014). [PubMed: 25038413]
31. T Gillespie D A General Method of Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. vol. 22 (1976).
32. Haghverdi L, Büttner M, Wolf FA, Buettner F & Theis FJ Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods 13, 845–848 (2016). [PubMed: 27571553]
33. Hoppe PS et al. Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. Nature 535, 299–302 (2016). [PubMed: 27411635]
34. Buggenthin F et al. Prospective identification of hematopoietic lineage choice by deep learning. Nat Methods 14, 403–406 (2017). [PubMed: 28218899]
35. Strasser MK et al. Lineage marker synchrony in hematopoietic genealogies refutes the PU.1/GATA1 toggle switch paradigm. Nat Commun 9, 2697 (2018). [PubMed: 30002371]

36. 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). [PubMed: 18371378]

37. Laslo P et al. Multilineage Transcriptional Priming and Determination of Alternate Hematopoietic Cell Fates. Cell 126, 755–766 (2006). [PubMed: 16923394]

38. Hormoz S et al. Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. Cell Syst 3, 419–433.e8 (2016). [PubMed: 27883889]

39. Loeffler D et al. Mouse and human HSPC immobilization in liquid culture by CD43- or CD44-antibody coating. Blood 131, 1425–1429 (2018). [PubMed: 29453290]

40. Manno GL et al. RNA velocity of single cells. Nature 560, 494 (2018). [PubMed: 30089906]

41. Hilsenbeck O et al. Software tools for single-cell tracking and quantification of cellular and molecular properties. Nat. Biotechnol 34, 703–706 (2016). [PubMed: 27404877]

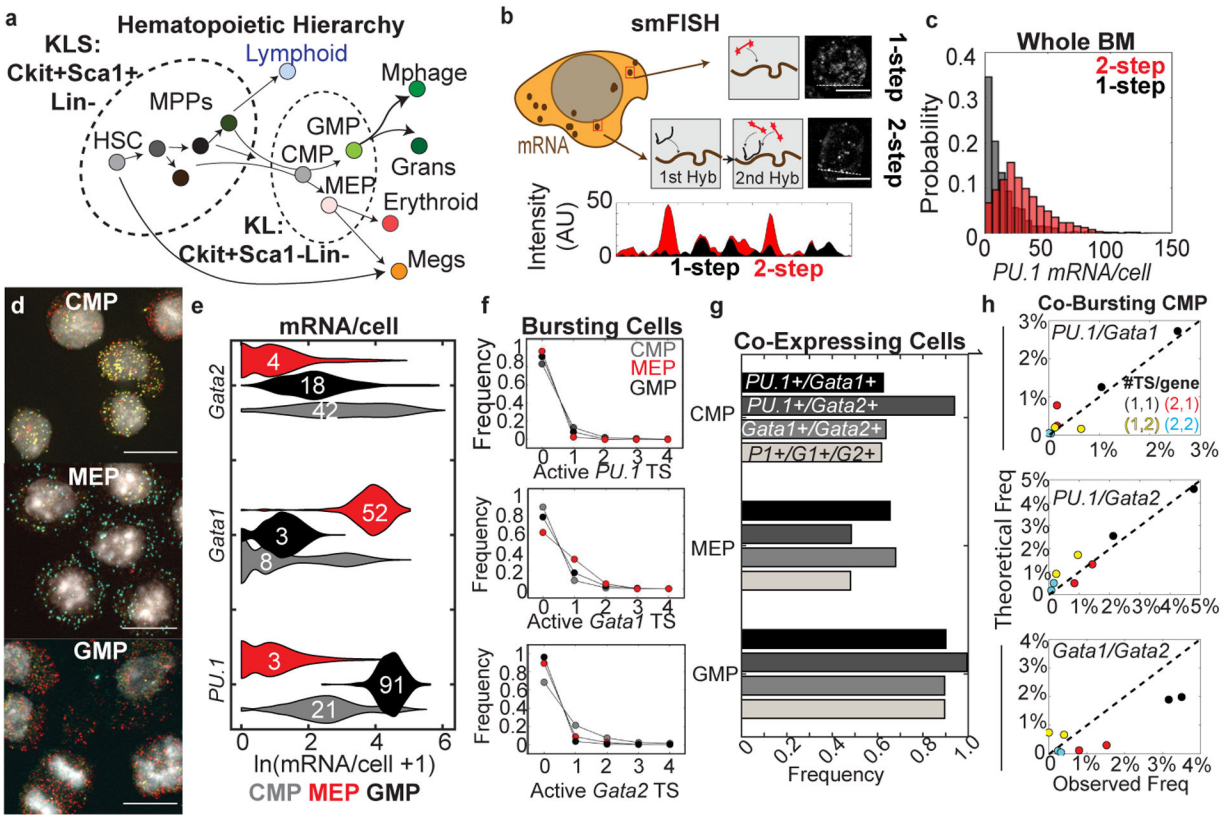


Fig. 1|. Stochastic Bursting of mRNAs Drives Co-expression of Antagonistic TF in HSPC.

a, Schematic of hematopoietic hierarchy. **b**, Description of smFISH using two-step hybridization method. Bottom panel are line plots of signal above background. **c**, Quantification of *PU.1* molecules per bone marrow mononuclear cell using 1-step or 2-step smFISH reaction. **d**, Filtered images of CMP, GMP, and MEP cells stained by smFISH for *PU.1* (Cy5, Red pseudocolor), *Gata1* (AlexaFluor 594, cyan pseudocolor), and *Gata2* (Cy3, yellow pseudocolor). Scale bars= 10μm, DNA in gray pseudocolor. **e**, Violin plots (Area normalized) of the natural log normalized (mRNA+1/cell) distribution for each gene. Overlaid numbers are the mean copy number/cell (CMP n=3174, GMP n=364, MEP n=1113). **f**, Burst frequency for each gene in each HSPC subpopulation. **g**, Frequency of cells co-expressing *PU.1/Gata1/2*. **h**, Comparison of observed co-bursting frequencies versus theoretical frequencies derived from statistical independence. Color indicates which combination of bursting patterns is being tested, e.g. (1,2) in the top panel means the frequency of cells with 1 active *PU.1* site and 2 active *Gata1* sites. Dashed line is y=x. Data in **(d-h)** are derived from 2 independent experiments for CMP and MEP and 1 experiment for GMP.

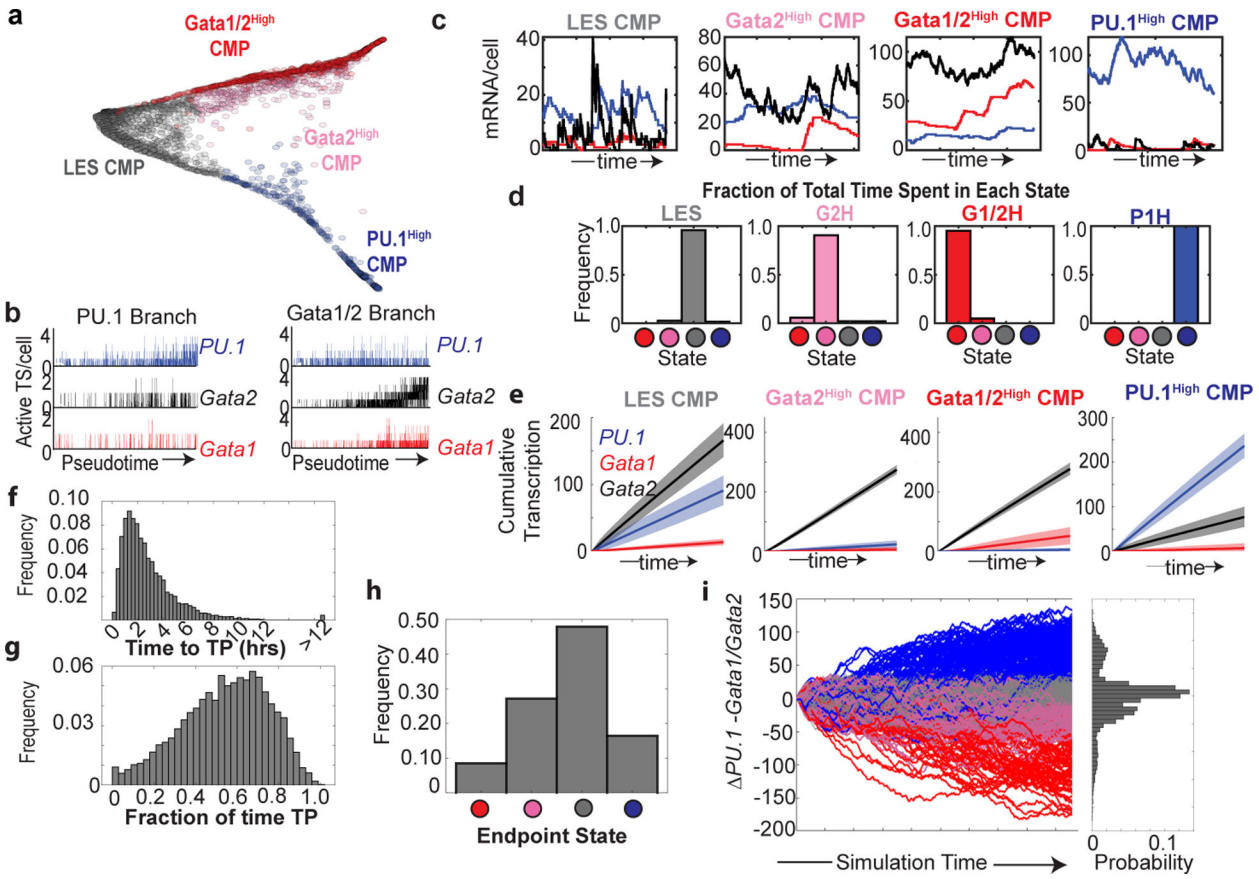


Fig. 2|. Inferred Dynamics of the *PU.1/Gata1/Gata2* Network in CMP.

a, Diffusion pseudo-time mapping of CMP cells, colored 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 cumulant nascent mRNA produced during the simulation. Line indicates mean among simulations; shaded regions are +/− standard deviation. n = 10,000. **f-g**, Time dependent behavior of simulated cells in the LES parameter regime, initialized at 0 mRNAs for all three genes at t = 0. **(f)** Histogram and time from start of simulation to first time point of instantaneous co-expression, i.e. triple positive or “TP”. All first TP events >12 hours were pooled together. **(g)** Histogram of total simulation time spent in TP (mean =56.8%, std=20.6%, n=10,000). **h-i**, Analysis of noise-derived transitions between states and efficacy of system evolution from LES. **(h)** Frequency each endpoint state after 12-hours of simulation time, initialized in the LES state (n=10,000). **(i)** Behavior of simulation trajectories over time. Colored based on endpoint state. Right is marginal distribution of endpoints.

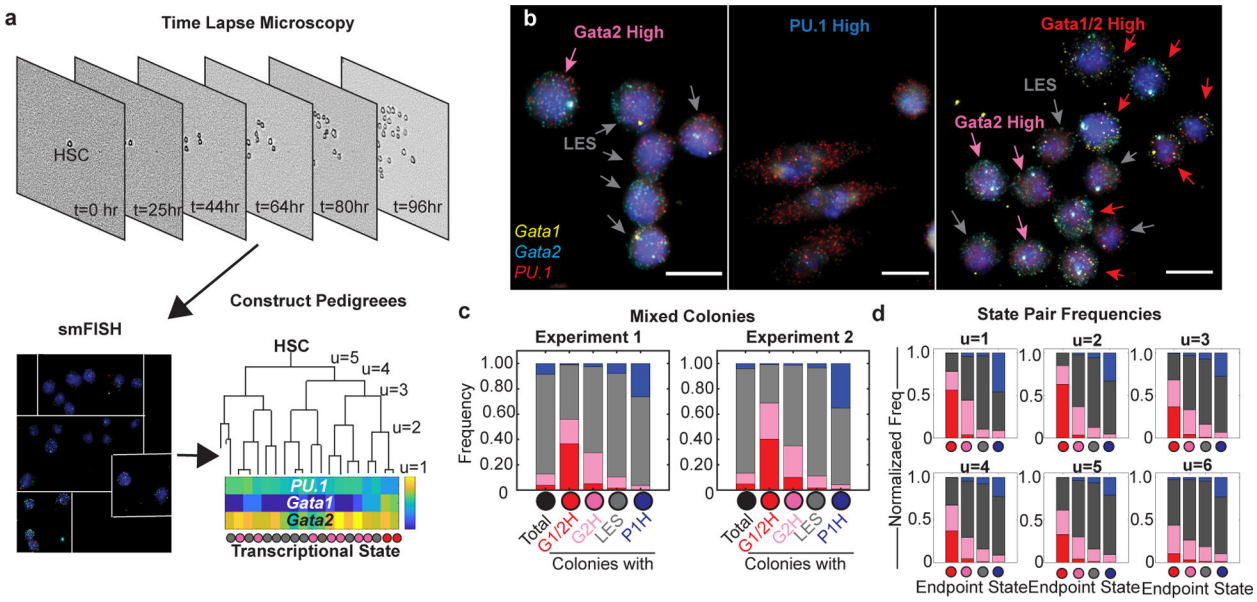


Fig. 3|. Transcription State Correlation Among Clonal Progeny of Single HSC.

a, Schematic of experimental workflow. smFISH image is stitched composite of 4 separate fields of view. Heatmap associated with the pedigree represents the ln(mRNA+1/cell). Colored 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 bar = 10μm. **c**, Frequency of states within mixed colonies conditional on the presence of each state. Total represents the frequency of states in all cells analyzed at the endpoint. The empiric distribution of the 4 HSPC states at the 96-hour endpoint was 2.9% (G1/2H), 14.50% (G2H), 6.9% (P1H), and 74.8% (LES) (Experiment 1 = 33 colonies. Experiment 2 = 87 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 colored circles under each bar plot.

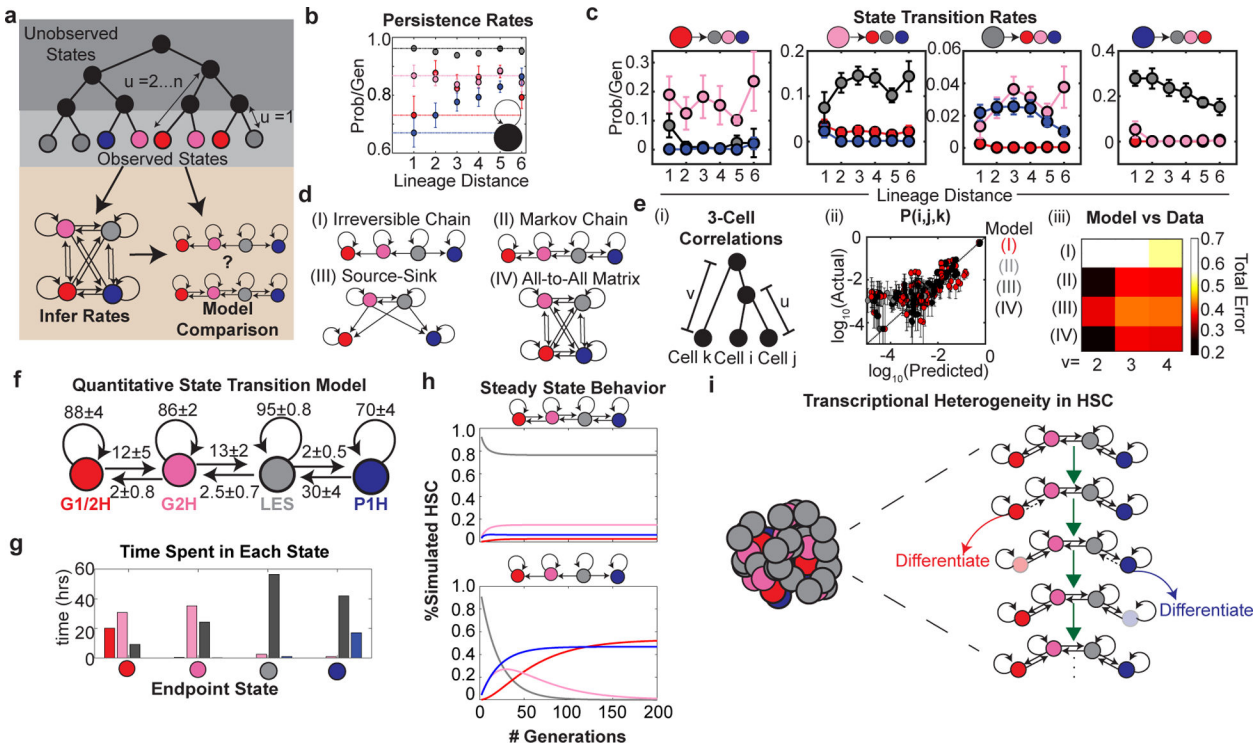


Fig. 4|. Stochastic and Reversible HSC Transcription State Dynamics.

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)** Schema of tested state transition models **(e)** (i) Schematic of three-point state frequencies. (ii) 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). (iii) Total error between theory and observed frequencies at v =2:4 for each model. Models with irreversible edges between states have higher error, i.e. less predictive value, than those with reversible edges. **f**, Average +/− standard error transition probabilities 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 colors correspond to each state as in (c). **i**, Proposed model of reversible transcription state transitions connecting *PU1-Gata* states in early HSPC.