Supplementary Information

Transcriptome-wide noise controls lineage choice in mammalian progenitor cells

Hannah H. Chang[1,2,3], Martin Hemberg[4*], Mauricio Barahona[4], Donald E. Ingber[1], and Sui Huang[1†]

[1]*Vascular Biology Program, Department of Pathology and Surgery, Children's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA*
[2]*Program in Biophysics, Harvard University, Boston, Massachusetts 02115, USA*
[3]*MD-PhD Program, Harvard Medical School, Boston, Massachusetts 02115, USA*
[4]*Department of Bioengineering and Institute for Mathematical Sciences, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom*

[*] *Present address: Department of Ophthalmology, Children's Hospital Boston, Boston, Massachusetts 02215, USA.*
[†] *Present address: Institute for Biocomplexity and Informatics, University of Calgary, Calgary, Alberta T2N 1N4, Canada*

Contents

S1. Supplementary Methods

Cell cycle arrest and cell cycle analysis

Cells were treated with a 24 h pulse of 40 μM Lovastatin (Sigma-Aldrich) to deplete cells in S phase[1]. To analysize cell cycle status, a combined BrdU-incorporation and PI staining protocol was used, following manufacturer's instruction. Briefly, cells were pulsed for 1 h with 10 μM BrdU (Sigma-Aldrich) and then fixed with cold 70% v/v ethanol. After washing with PBS + 0.5% BSA, cells were denatured with 2M HCl for 20 min, neutralized with 0.1 M sodium borate, then labeled with anti-BrdU-FITC monoclonal antibody (Becton Dickinson) and 10ug/ml Propidium Iodide stain, and analyzed by flow cytometry. Analysis of cell cycle status in live cells was performed with Hoerchst 33342 (Invitrogen) stain at 5μg/ml final concentration and analyzed with a UV-laser equipped flow cytometer. Cell cycle data was then analyzed using the FlowJo 2.2.2. software package (Tree Star) to determine the relative proportions of cells in G_0/G_1, S, and G_2/M cell cycles.

Quantitative real-time reverse transcription (RT) PCR

Total RNA was isolated from 8 – 20 x 10^6 cells by using the RNeasy Mini RNA isolation kit (Qiagen). RNA was reverse-transcribed with Omniscript RT-PCR kit (Quiagen) in accordance with the manufacturer's protocol and used to test primer activity. Real-time PCR was performed on ~250 ng of total RNA/sample with the QuantiTect SYBR Green PCR kit (Qiagen) in accordance with the manufacturer's instructions. Amplification conditions were as follows: 40 cycles of denaturation at 94 °C for 15 s, annealing at 55°C for 30 s, and extension at 72 °C for 30 s using the Mx4000 (Stratagene) or 7300 (Applied Biosystems) realtime-PCR machines (Stratagene). Primers for GATA1 were (Right: CAGGGCAGAATCCACAAACT, Left: TCCTCTGCATCAACAAGCC), Sca-1 (Right: GGTTCTTTAGGCTGGCAGTG, Left: GGGAAGTTTCCATGGTGAAG) (from the qPrimerDepot database http://mouseprimerdepot.nci.nih.gov/), PU.1 (Right: TGACTACTACTCCTTCGTGG, Left: GATAAGGGAAGCACATCCGG), and GAPDH (right: ACCACAGTCCATGCCATCAC, Left: TCCACCACCCTGTTGCTGTA). Specificity was verified by melt-curve analysis and agarose-gel electrophoresis. Results are standardized for GAPDH expression levels and are expressed as fold induction compared with the levels (set to 1) detected in the sample with the lowest expression.

Western Blot analysis

1 – 5 x 10^6 cells were pelleted and homogenized with the appropriate volume of RIPA buffer (Boston BioProducts) containing 50mM Tris-HCl, 150 mM Nacl, 1% NP-40, 0.5% Sodium deoxycholat, and 0.1% SDS, and protease-inhibitor cocktail (Roche) and sheared with multiple passages through a syringe. After measurement of protein yield using the D_c Protein Assay (Bio-Rad), whole-cell lysates were boiled for 5 min at 95°C with 20% sample-loading buffer. 30-40 μg of total cell lysate were subjected to electrophoresis on 4-20% SDS-polyacrylamide gradient gels (Bio-Rad) and transferred to nitrocellulose membranes. Following blocking with 5% milk/PBST (phosphate buffered saline with 0.1% Tween 20), the membrane was probed either with a 1:200 dilution of anti GATA1-N6 antibody (Santa Cruz sc-265) or a 1:1000 dilution of anti PU.1 antibody (Santa Cruz sc-352). Antibody binding was detected with a 1:10000 dilution of peroxidase labeled

anti-rat IgG (Santa Cruz) or anti-rabbit IgG (Vector) and luminescence was detected with Supersignal West Dura Signal reagents (Pierce).

S2. Supplementary Discussion

1. What other factors could contribute to the observed level of heterogeneity in Sca-1 within one clonal population (Fig. 1 in the main text)?

Before studying clonal heterogeneity as an intrinsic phenomenon with potential biological function, we experimentally considered the following possible (trivial) sources for the observed variability:

(1) Measurement noise (flow cytometry): the upper bound of the error due to fluctuations in the measurement process (e.g., machine noise) is given by the spread of the signal obtained from standardized MESF[2] beads that have uniform amount of fluorescence on each bead (within manufacturing error). This error was 2-fold (Fig. 1b in the main text).

(2) Cell size and cell cycle: Because absolute gene expression levels are affected by cell size[3], we examined whether the observed variability in Sca-1 levels reflects cell size variations. The projected area of clonal EML cells ranged over 1.5-fold, which was associated with a 1.7-fold difference in mean Sca-1 expression (Supplementary Fig.1a). Taken together, this would only account for less than 1% of the total Sca-1 heterogeneity we observed. Since some proteins exhibit cell cycle dependence even without an explicit role in cell division[4], cell cycle asynchrony in populations of clonal cells could also be a source of Sca-1 clonal heterogeneity. However, clonal cells in the G0/G1 and G2/M cell phases independently showed greater than 500-fold range in Sca-1 expression, differing less than 2-fold in mean Sca-1 expression (Supplementary Fig.1b), again a result that cannot explain the observed Sca-1 heterogeneity.

Thus, the variation in Sca-1 expression in clonal EML cells cannot be trivially attributed to measurement noise, variation of cell size, or asynchrony in cell cycle.

2. What biological process may drive the (re)generation of the parental Sca-1 distribution from the three sorted, more homogeneous population fractions?

Here we present experiments or arguments suggesting that some commonly assumed mechanisms are unlikely to provide complete explanations for the dispersion of Sca-1 and the slow relaxation to the parental distribution, although their partial contribution cannot be excluded.

(1) Shift of cell population demography by overgrowth of a subfraction of cells with the appropriate Sca-1 expression level: The sorted fractions could be contaminated by a few residual cells from other fractions that outgrow more slowly dividing cells, thereby restoring the missing populations. Because the Sca-1^{Low} outlier fraction must accumulate cells with higher Sca-1 levels to reconstitute the original histogram (and accordingly, the Sca-1^{High} fraction must accumulate low expressors), Sca-1^{Mid} cells would be the common "contaminant" that could override both outlier fractions. However, the growth rate of the Sca-1^{Mid} fraction was not higher than that of the outlier fractions (Supplementary Fig. 3). Moreover, greater than 98% purity was obtained for all sorted fractions as obtained by reanalysis of sorted samples.

(2) Mutations affecting Sca-1 expression distribution. Genetic mutations may in principle be responsible for the heterogeneity and slow changes in Sca-1 expression level

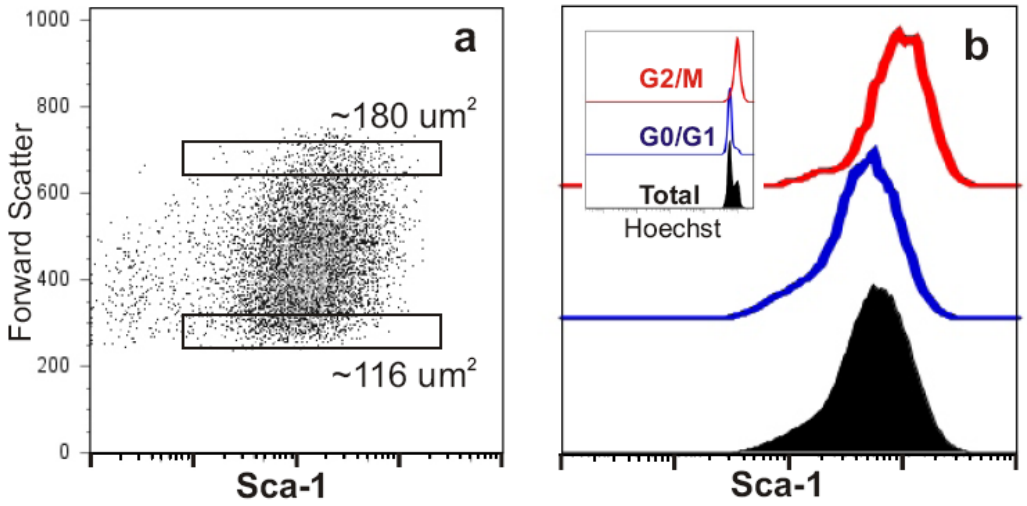
distribution. However, this is unlikely since thousands of diverse mutations would have to be generated within less than 9 days (~ 12 cell divisions), each of which would have to confer both a robust growth advantage and a distinct, stable level of Sca-1 surface expression to collectively cover a greater than 10-fold range of population variability.

(3) "Gene expression noise" as basis for dispersion. Propagation of noise from gene transcription[5] to the protein level as a source of clonal heterogeneity can be ruled out because there was no statistically significant difference in the Sca-1 mRNA levels of the Sca-1^{Low}, Sca-1^{Mid}, and Sca-1^{High} fractions, as determined by real-time PCR (Supplementary Fig. 4). However, random fluctuations at later stages in Sca-1 surface expression (translation, membrane localization via GPI anchor, trafficking) may play a role.

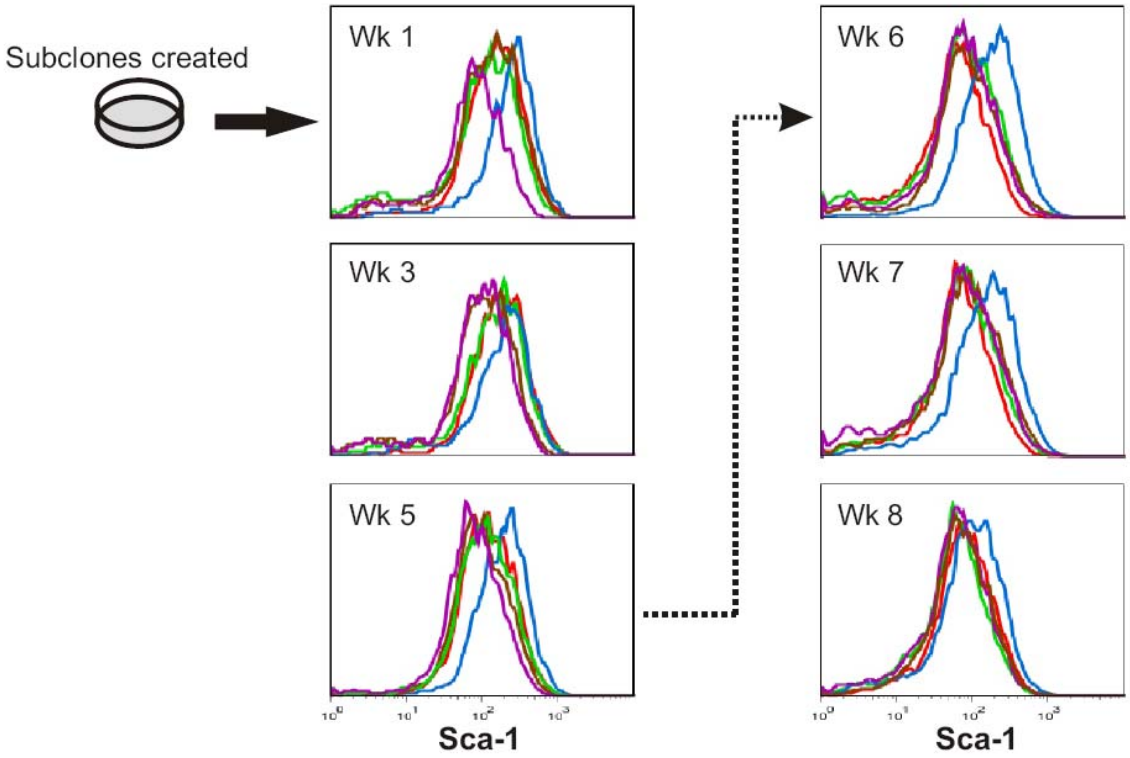
(4) Uneven partitioning of Sca-1 proteins in cell division. The uneven distribution of cellular molecules to the daughter cells during cell division has long been suggested to be a mechanism that generates population heterogeneity[6]. Here we do not explicitly study this mechanism. However, we observed that the width (spread) of the histogram of Sca-1 expression levels increased significantly within 24 hr. The rate of cell division is much slower (Supplementary Fig. 3a) such that only a fraction of cells would have undergone cell division in 24 hrs. Thus, while uneven partitioning may still be a source of Sca-1 heterogeneity in long-term cell cultures, it is unlikely to be the sole driving force behind the restoration of the parental distribution from a narrow distribution.

While here we do not establish the driving force for the diversification of Sca-1 levels and the mechanisms that slow the underlying kinetic, it is possible that they result from the joint effect of several processes including the ones discussed above.

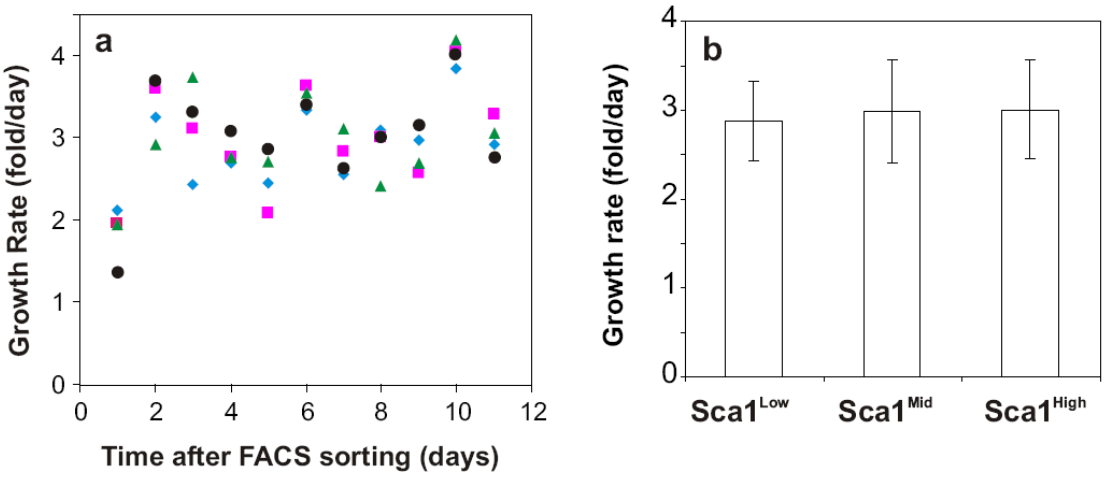
S3. Supplementary Figures



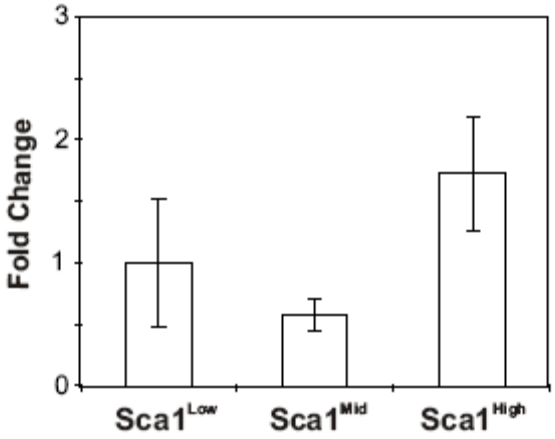
Supplementary Figure 1. Robust clonal heterogeneity. a, Weak correlation between cellular Sca-1 expression and cell size (projection area) revealed by Fluorescence Intensity – Forward Scatter dot plot. **b,** Clonal cells in G0/G1 (blue), G2/M (red) and combined cell cycle phases (black), distinguished by Hoechst stain (inset) showed minor differences in overall range and mean Sca-1 expression.



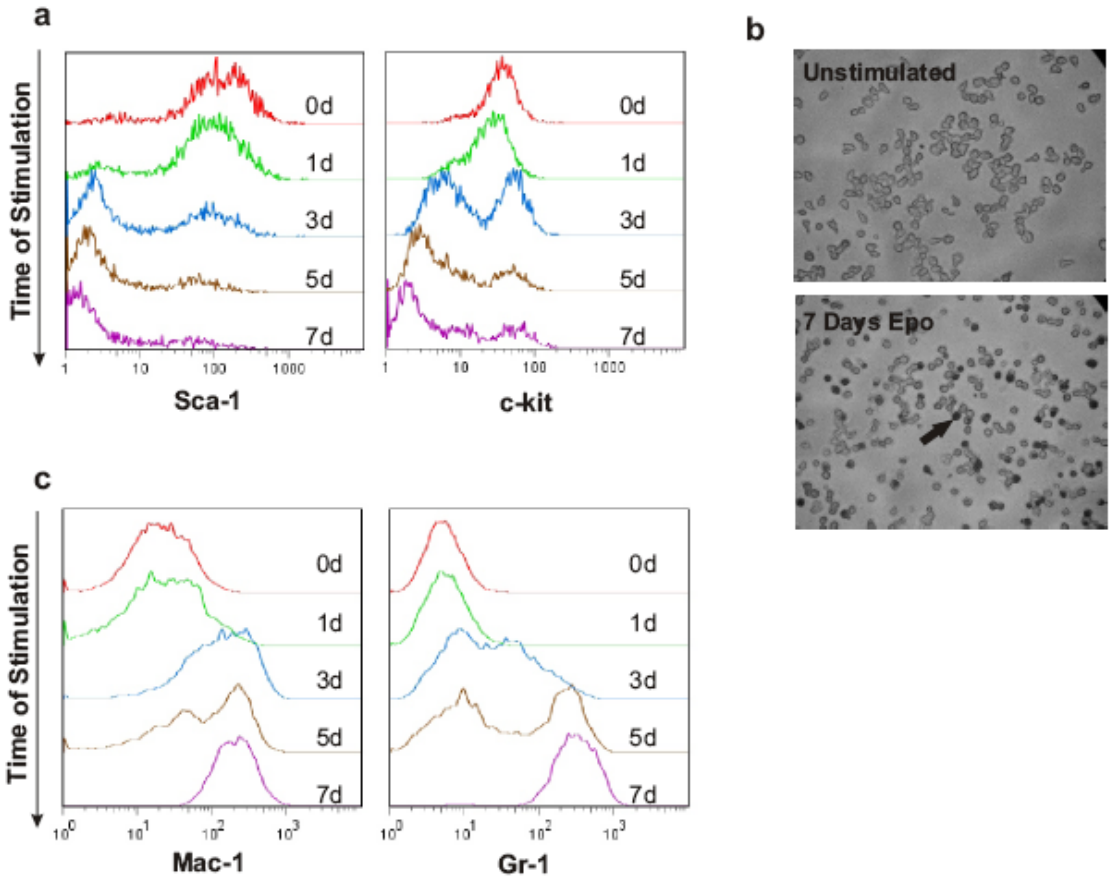
Supplementary Figure 2. Clonal heterogeneity in Sca-1 expression among single-cell-derived subclones converged towards that of the original parental clone. Population distribution of non-stimulated, baseline Sca-1 expression for four single-cell derived subclones (purple, brown, green, blue solid lines) represented by flow cytometry histograms exhibited convergence towards the parental clone histogram (red) over eight weeks (Wk) in normal growth culture despite one cycle of freeze/thaw after week 5.



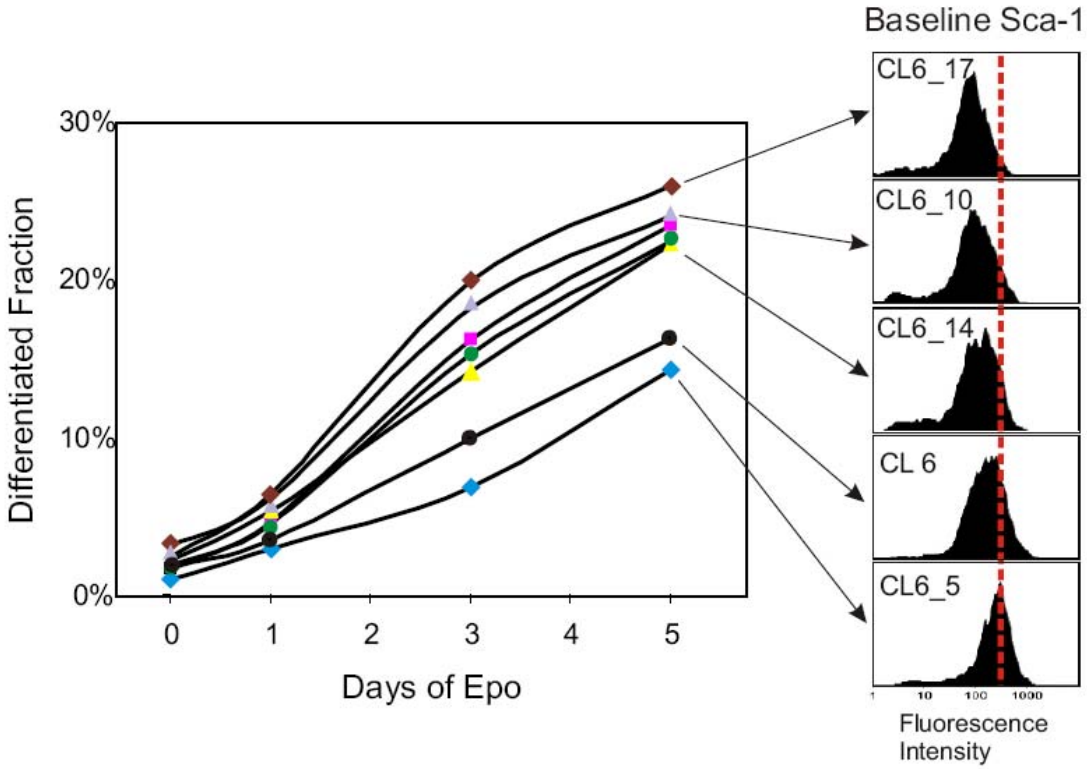
Supplementary Figure 3. Growth rates of sorted fractions. a, The growth rates of the Sca-1^{Low} (blue diamonds), Sca-1^{Mid} (magenta squares), Sca-1^{High} (green triangles) sorted fractions, and a mock-sorted control (black circles) were calculated as the fold difference between two daily measurements. **b,** The three sorted fractions had comparable growth rates overall, as shown by the mean and standard deviation of growth rates over all 11 times points in **a**.



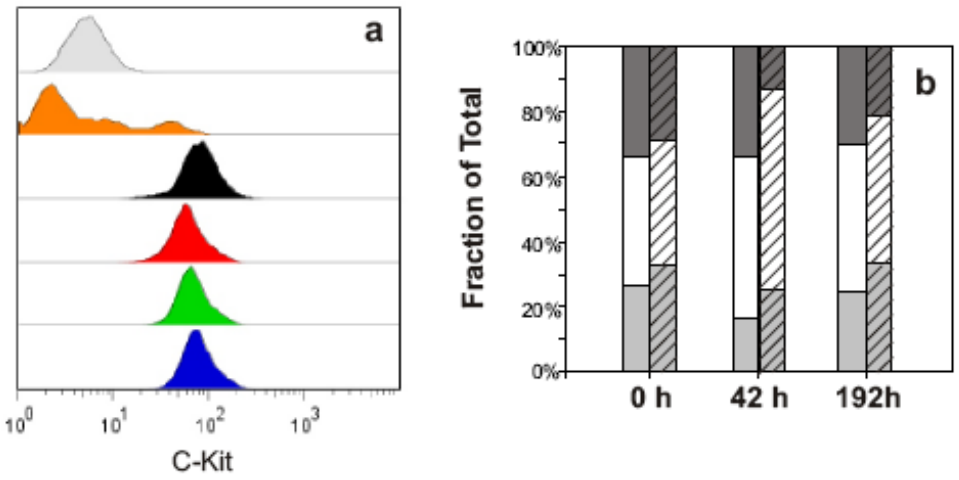
Supplementary Figure 4. Sca-1 mRNA levels in sorted fractions. Sca-1 mRNA levels in the Sca-1^{Low}, Sca-1^{Mid}, and Sca-1^{High} fractions analyzed by quantitative RT-PCR did not differ significantly. Results represent the mean and standard errors from quadruplicate measurements. Each value has been standardized for GAPDH expression levels and is expressed as fold induction compared with the levels (set to 1) detected in the Sca-1^{Low} sample. (Sca-1^{Low} vs. Sca-1^{Mid}, p-value > 0.4, Sca-1^{Mid} vs. Sca-1^{High}, p-value > 0.5 by Student's t-test.)



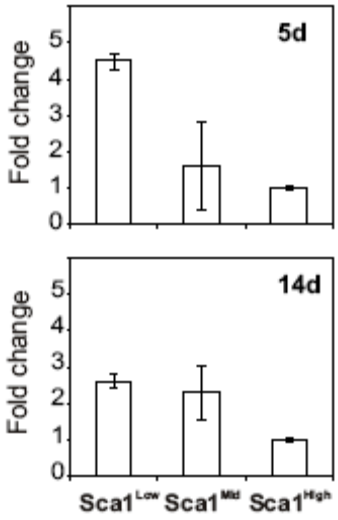
Supplementary Figure 5. Differentiation of EML cells into pro-erythrocytes and myelocytes. a, EML cells are positive for the stem-cell markers c-kit and Sca-1 without stimulation ('0d') as monitored by flow cytometry. Cells tracked over seven days (d) of Epo treatment showed loss of Sca-1 and c-kit expressions ('1d', '3d', '5d', and '7d'). **b,** Seven days of Epo stimulation ('7 Day Epo') results in positive benzidine staining (black arrow) as compared to the absence of staining in the un-stimulated cells ('Unstimulated'). Phase contrast images taken at 20x. c, EML cells stimulated with IL-3 and GM-CSF showed gain of expression for the myeloid-lineage specific markers Mac-1 and Gr-1 within seven days.



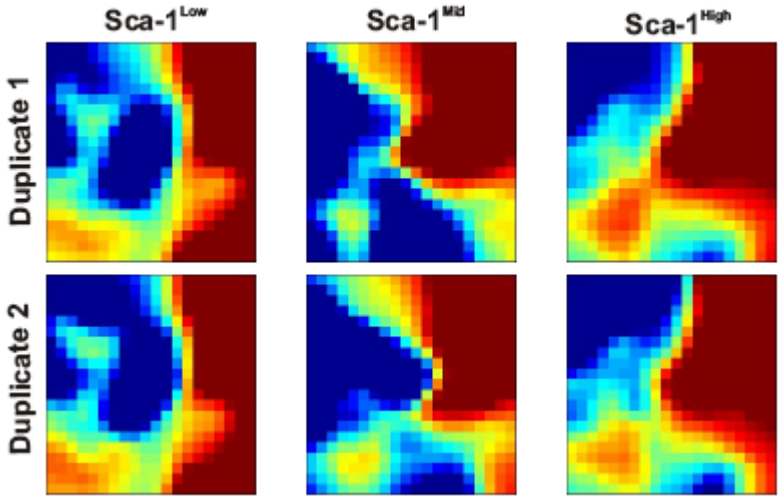
Supplementary Figure 6. Sca-1 clonal heterogeneity governs differentiation potential among individual subclones. Mean baseline Sca-1 expressions (histograms on the right) for four representative subclones (CL6_17, CL6_10, CL6_14, CL6_5) and the parental clone (CL 6) were inversely proportional to the rate of commitment to pro-erythrocytes upon stimulation with Epo (left). Subclones were generated by expansion of randomly-selected cells from the parental population. Burgundy diamonds, CL6_17; grey triangles, CL6_10; magenta squares, CL6_15; green circles, CL6_20; yellow triangles, CL6_14; black circles, CL6 and blue diamonds, CL6_5. Rank-order of differentiation kinetics for individual subclones is preserved across all four time points, $p < 10^{-6}$ (permutation test).



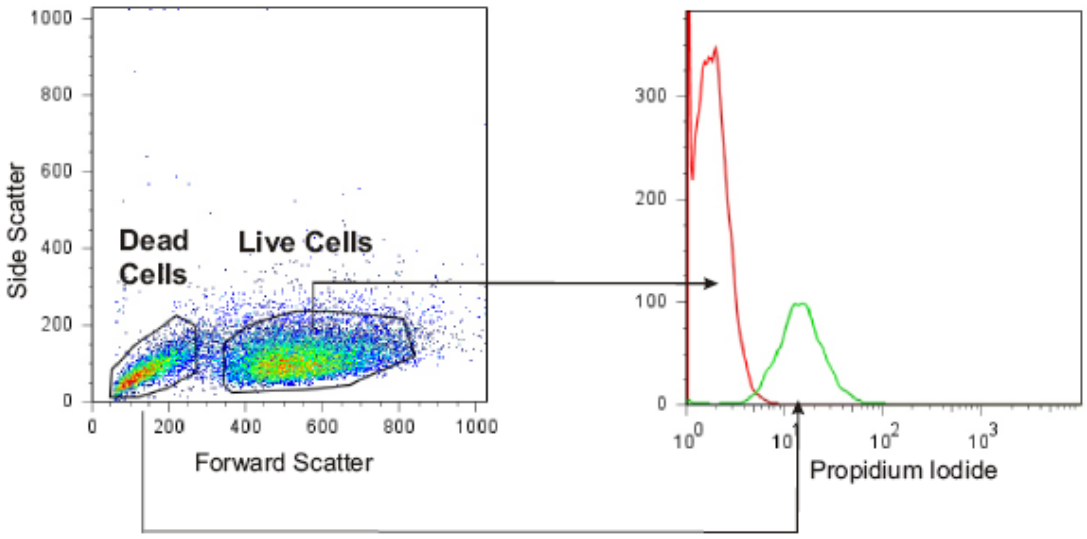
Supplementary Figure 7. Sca-1^{Low} cells are not spontaneously differentiated pro-erythrocytes. a, Sca-1^{Low} cells (red) showed positive c-kit expression compared to differentiated pro-erythocytes (orange) and isotype control (light grey) as do the Sca-1^{Mid} (green), Sca-1^{High} (blue), and parental populations (black). **b,** Distribution of cells in the G0/G1 (light grey), S (white), and G2/M (dark grey) cell cycle phases among the Sca-1^{Low} fraction (solid bars) and a mock-sorted whole population control (dashed bars) are similar at 0, 42, and 192 hours (h) after FACS isolation.



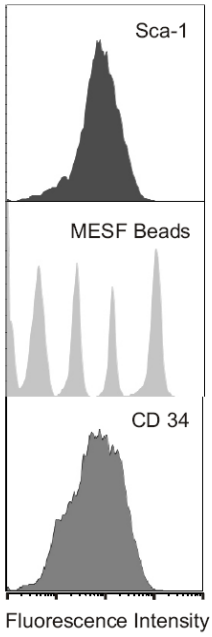
Supplementary Figure 8. GATA1 mRNA expression among sorted Sca-1 fractions. Quantitative RT-PCR analysis of GATA1 mRNA levels in Sca-1 sorted fractions after 5 or 14 days (d) of regular culture. Means ± standard error of triplicates shown.



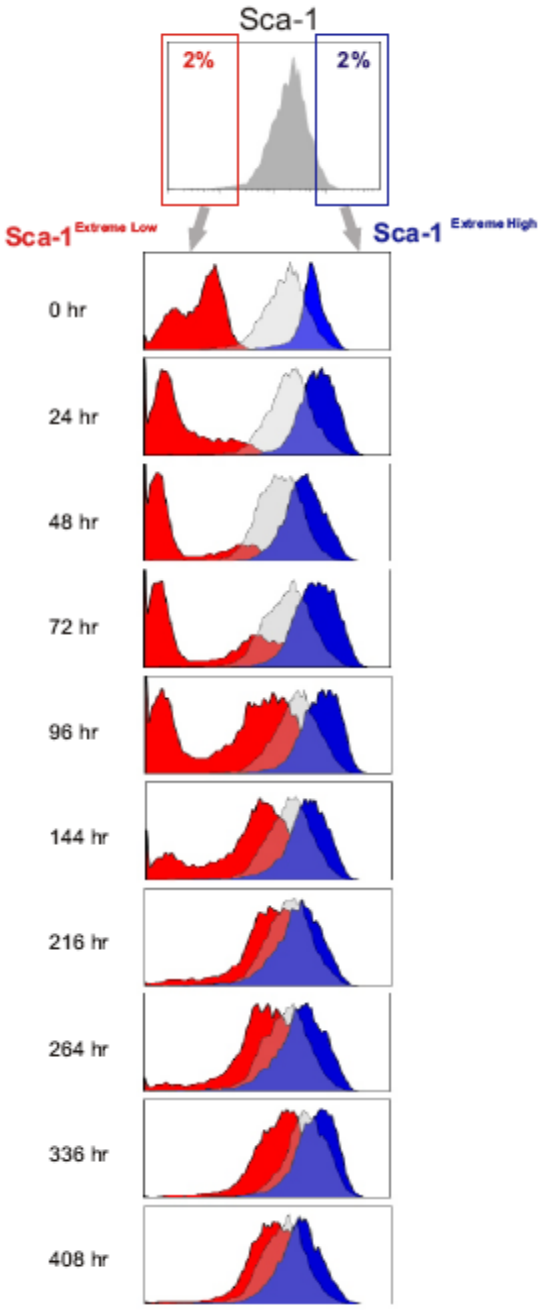
Supplementary Figure 9. Global gene expression analysis showed high duplicate accuracy. Hybridization duplicates for global gene expression analysis using Illumina microbead chips for the Sca-1^{Low}, Sca-1^{Mid}, and Sca-1^{High} cell fractions showed nearly identical GEDI maps, while the transcriptome dissimilarities between the cell fractions was recapitulated. Pearson's correlation coefficient were > 99.0% for all three pairs of duplicates.



Supplementary Figure 10. Live versus dead cells. Live EML cells showed negative propidium iodide (PI) staining and high forward scatter (FSC) whereas dead cells show positive PI-staining and low FSC. The accurate correlation between FSC and PI-staining was used to remove dead cells from all analysis of flow cytometry data by gating out cells with low FSC.



Supplementary Figure 11. An additional example of clonal heterogeneity. Heterogeneity in expression of the hematopoietic progenitor cell surface protein CD 34 (bottom) within clonal cells is comparable to that of Sca-1 (top), and much larger than the resolution limit of flow cytometry approximated by measurements of reference MESF beads (middle).



Supplementary Figure 12. Additional analysis: Restoration of heterogeneity from sorted "extreme Sca-1 expressors". Clonal cells with the highest (Sca-1$^{\text{Extreme High}}$ in blue) and lowest (Sca-1$^{\text{Extreme Low}}$ in red) 2% Sca-1 expression (rather than 15% as in the main text) also re-established the parental extent of clonal heterogeneity (grey) in separate cultures. However, the rate of restoration was very slow and the process incomplete even after 408 hrs. Interestingly, a spontaneously differentiating subpopulation is observed among the Sca-1$^{\text{Extreme Low}}$ fraction, which generated a new population of Sca-1$^{\text{neg}}$ cells. This gave rise to the familiar bimodal distribution indicative of fate commitment through a discontinuous, "all-or-none" switching process (see reference Chang et al in the main text). Notably, cells that did not spontaneously differentiate in the Sca-1$^{\text{Extreme Low}}$ fraction were capable of regenerating the parental distribution, but with a very slow rate.

S4. Supplementary Table

	Low-0d	Mid-0d	High-0d	Low-6d	Mid-6d	High-6d	Epo-7d
Low-0d		0.0266 ± 0.006	0.0548 ± 0.011				0.0791 ± 0.0215
Mid-0d			0.0614 ± 0.010				0.0844 ± 0.0267
High-0d							0.1578 ± 0.0323
Low-6d					0.0095 ± 0.002	0.0066 ± 0.0018	
Mid-6d						0.0116 ± 0.0023	
High-6d							
Epo-7d							

Supplementary Table 1. Quantified dissimilarity of global gene expression between samples. Pair-wise dissimilarity between the Sca-1^{Low} (Low), Sca-1^{Mid} (Mid), Sca-1^{High} (High) samples at 0 and 6 days (d) and a terminally-differentiated control sample (Epo-7d) were calculated based on the normalized gene expression levels for 2997 filtered genes with $1 - R$ where R is the Pearson's correlation coefficient which ranges from 0 to 1, with 0 being the most similar and 1 being the most different (Methods). Bootstrapping was performed by randomly selecting 30% of the genes in any sample to calculate the pair-wise dissimilarity metric and repeating the procedure 10,000 times to generate the standard deviations reported above.

S5. Theoretical Methods

As motivated in the main text, the temporal evolution of the distribution of cellular Sca-1 abundance (log fluorescence intensity values) cannot be fitted to a model consisting of a single Gaussian. Single-Gaussian fits for the Sca-1 distributions at the stationary time points were poor ($p<10^{-50}$, Kolmogorov-Smirnov test). This is also evident from the hump in the long left tail occurring at different time points, a signature of multimodality. These features suggest that the restoration of the parental distribution is not a simple noise-driven, mean-reverting, equilibrium-seeking process in a smooth potential, such as an Ornstein-Uhlenbeck process. Taken together, this led to the hypothesis that the multi-modal character stems from the fact that the cell population is discretely heterogeneous, consisting of two (or more) distinct but overlapping subpopulations. Such a behavior may result from complex regulatory processes that involve multi-stability.

The purpose of the modeling and analysis presented below is to corroborate the notion of multiple subpopulations with respect to Sca-1 steady-state expression by testing whether the data can be better fitted to a multi-rather than single Gaussian distribution, without making assumptions concerning the unknown underlying molecular circuitries. The results show that a two-Gaussian model best fits the observed histogram evolution, and that the restoration of the parental distribution was predominantly driven by state transitions between the subpopulations.

S5.A. Fitting of fluorescence histograms

The experiments show that the stationary distribution of the log-fluorescence intensity value of the cell population presents multimodal features. Moreover, after sufficient time, the stationary distribution is reconstituted from all three sorted fractions: Sca-1^{High}, Sca-1^{Mid}, and Sca-1^{Low}. Therefore, the stationary distribution can be used to determine basic parameter values for a model of the distributions.

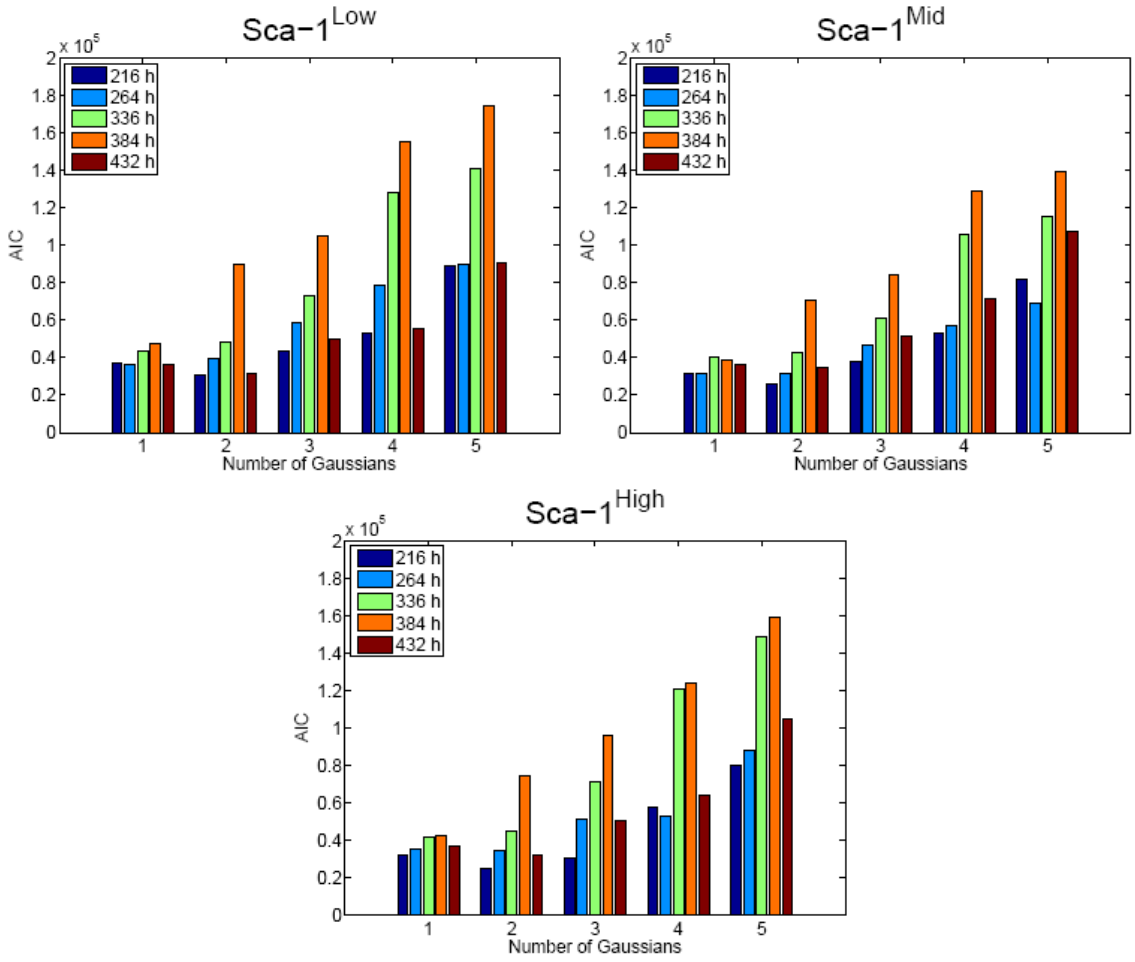
As a first approximation, the multimodal character of the data can be captured by a linear combination of *n* Gaussian distributions with different means and variances. The underlying assumption is that there are *n* subpopulations, where each subpopulation in isolation has a log-normal fluorescence distribution with different mean and variance. This leads to a *Gaussian mixture model* (GMM) with probability density function (PDF) given by

$$P(x) = \sum_{i=1}^{n} w_i \phi\left(x, \mu_i, \sigma_i\right) \qquad (1)$$

where ϕ is the density of the ordinary Gaussian distribution with mean μ_i and standard deviation σ_i and w_i is the *weight* of the ith component (subpopulation). The weights must satisfy the constraints $\sum_{i=1}^{n} w_i = 1$ and $w_i \geq 0$.

At long times, the histograms of the three sorted fractions converge and the parental population is reconstituted. To represent the observed multimodal stationary distribution we use a GMM, which is obtained by using the *expectation-maximization* (EM) algorithm[7] to fit a total of 15 histograms corresponding to the last five time points

of each of the three time course experiments (Sca-1^{High}, Sca-1^{Mid}, and Sca-1^{Low}).



Supplementary Figure 13. Number of components in GMM: selection through the Akaike Information Criterion (2)**.** The AIC was calculated from the likelihood of the fits provided by the EM-algorithm and is minimal for $n = 2$ in most cases (Sca-1^{Low} at t = 216h and t = 432h; Sca-1^{Mid} at t = 216h, 264h and 432h; Sca-1^{High} at t = 216h, 264h and 432h). The application of another model selection criterion[8] also selects $n = 2$ in almost all cases. In some cases, AIC is higher for n =2 than n =1 because the fluctuations make the bimodality less apparent.

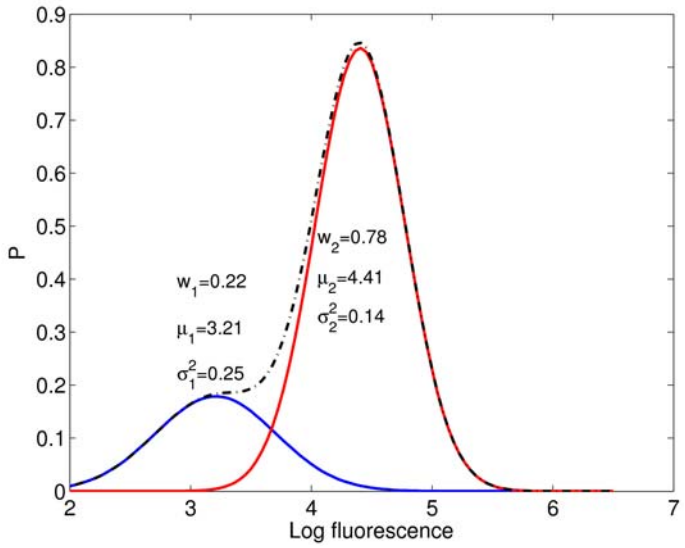
The number of Gaussians, n, to be fitted to the data is a user-specified parameter in the EM algorithm. To compare GMM's with different choices of n, we use *Akaike's information criterion* (AIC)[9,10]

$$AIC = -2\log L + 2n_p \qquad (2)$$

where L is the maximum likelihood of the model, n_p is the number of independently adjusted parameters within the model. When comparing different models, the one which minimizes Equation (2) provides the best combination of descriptive power and

parsimony. Supplementary Fig. 13 confirms that the GMM with $n = 2$ is the best choice based on the AIC. A similar model-selection framework[8], which integrates the EM algorithm with an information criterion based on the Fisher information matrix, also came out in favor of the $n = 2$ model.

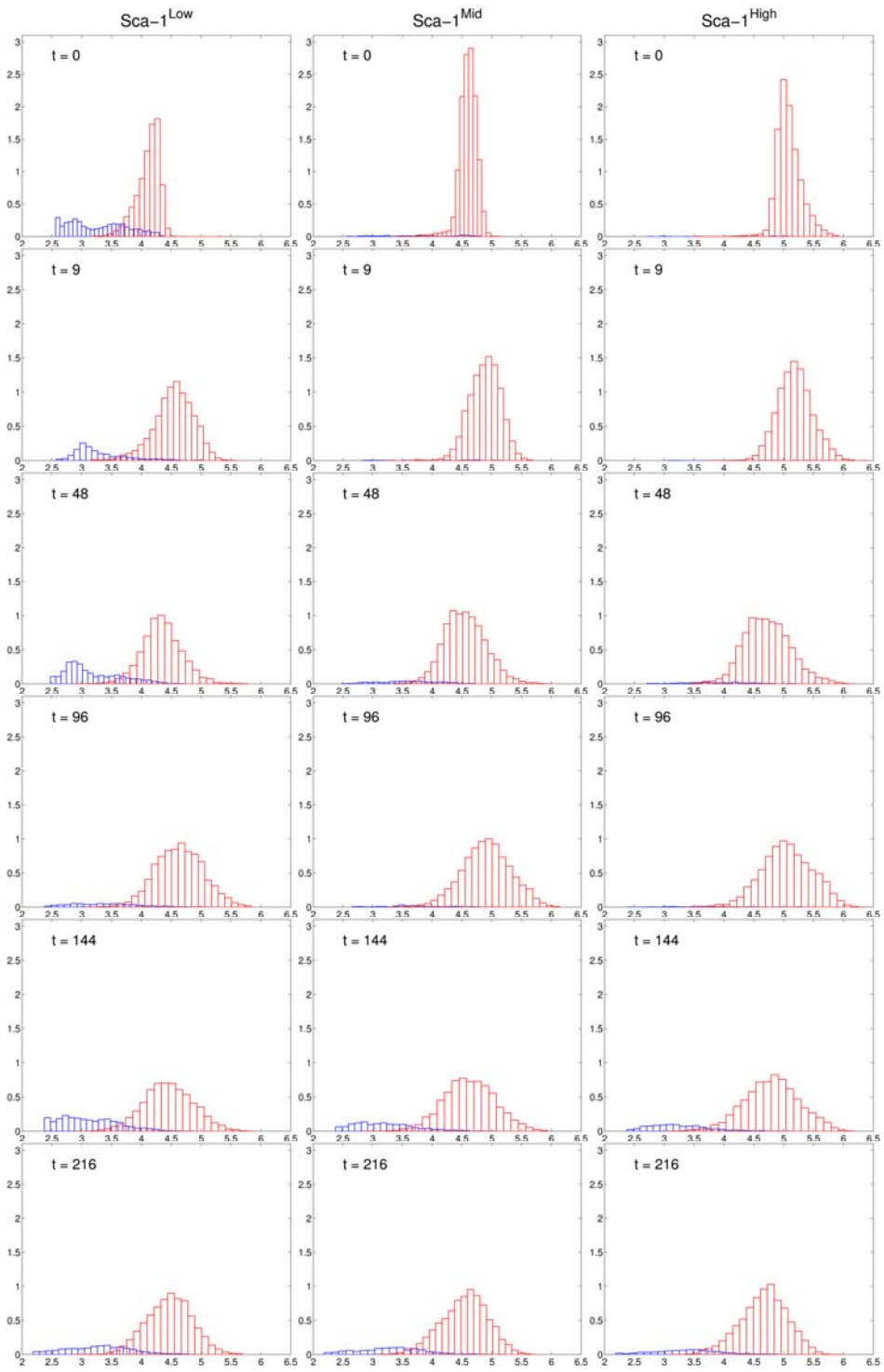
The average parameter values found for the GMM with two Gaussians are given in Supplementary Fig. 14. These parameter values are used for the partitioning of the cells in the populations of the individual measurement points.



Supplementary Figure 14. The GMM obtained from the stationary time points. The dash-dotted black line shows the PDF for the GMM and the solid lines for the two components G_1 (to the left in blue) and G_2 (to the right in red). The parameters for the two virtual subpopulations were obtained from the last five time points and are given in the figure. The numerical error of the EM algorithm for the parameter values is 10^{-3} and thus too small to be shown.

S5.B. Partitioning the fluorescence data based on the GMM

In order to describe the dynamics by which the parental population is reconstituted from the individual sorted experiments (Fig. 2a in main text) the *GMM binning algorithm* was developed. The objective of the algorithm is to partition the cells of each measured population into the two overlapping Gaussian distributions (Supplementary Fig. 14) that constitute the stationary GMM obtained in Section S5.A. The GMM binning algorithm sorts the $N(t)$ cells at time t according to their fluorescence value $c_j(t)$ by assigning a probability that each cell belongs to subpopulation $G_i : P_{ji} = \phi\left(c_j(t), \mu_i, \sigma_i\right)$. Each cell is then assigned randomly to the Gaussian $c_j(t)$ based on the normalized probability $P_{ji} / \sum_i P_{ji}$, as performed by the following pseudo-code:



Supplementary Figure 15. Results for the GMM binning algorithm for the data in Fig. 2a in the main text. From the top, partitions for t = 0, 9, 48, 96, 144, 216 h for the Sca-1^{Low} (left), Sca-1^{Mid} (center), and Sca-1^{High} (right) time course. For each panel, the sum of the red and blue histograms is equivalent to the data in Fig. 2a in the main text. The probabilistic nature of the binning algorithm means that the two inferred subpopulations overlap. For all panels, x-axis is the log fluorescence and the y-axis is cell number.

Algorithm 1 (GMM binning algorithm)

for each time *t* **do**

for each cell $c_j(t)$ **do**

$$P_{ji}(t) = \phi_i\left(c_j(t), \mu_i, \sigma_i\right), \quad i = 1,...n$$
$$G_i(t) \leftarrow \text{Assign}\left(P_{ji} / \sum_i P_{ji}\right)$$

end for

$$w_i(t) = N_i(t) / N(t)$$
$$\mu_i(t) = \sum c_{ji}(t) / N_i(t)$$
$$\sigma_i^2(t) = \sum\left(c_{ji}(t) - \mu_i(t)\right)^2 / \left(N_i(t) - 1\right)$$

end for

The function Assign uses a random to decide to which subpopulation G_i the sample cell $c_j(t)$ should be assigned. The weights can then be calculated as $w_i(t) = N_i(t) / N(t)$, where $N_i(t)$ is the number of cells assigned to subpopulation *i*. The mean, $\mu_i(t)$, and variance, $\sigma_i^2(t)$, for each subpopulation $G_i(t)$ can also be calculated. Because of the large number of cells in each sample, the algorithm is extremely robust and repeated runs give nearly identical results for the parameters (data not shown). The resulting histograms for the time points in Fig. 2a in the main text are shown in Supplementary Fig. 15. Due to the probabilistic nature of the binning algorithm, the two inferred subpopulations overlap.

S5.C. time evolution of the subpopulations

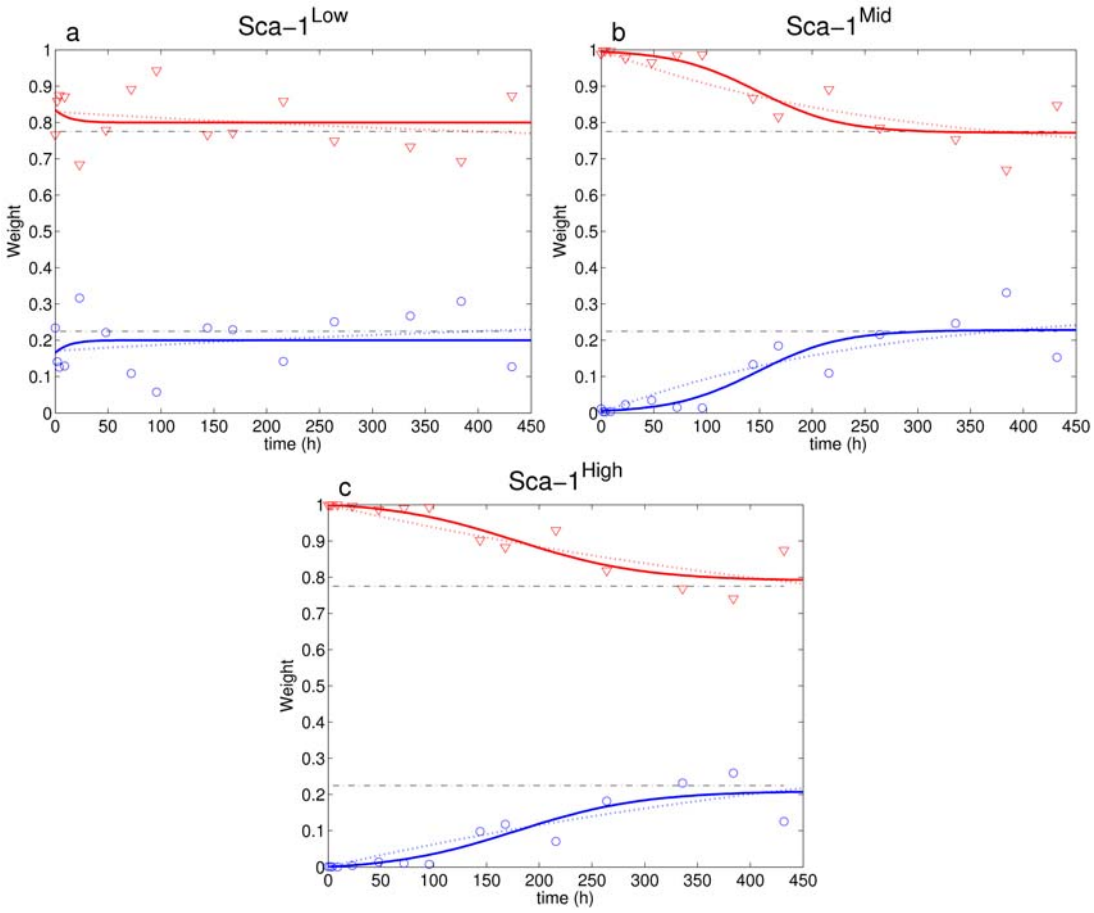
The separation of the fluorescence data into two subpopulations for each time point makes it possible to track the evolution of the relative weights w_i of the subpopulations, as shown in Supplementary Fig. 16. We have considered two models to describe the temporal evolution of the w_i, which we describe below.

(a) Lineage Model

A simple model of two interacting and growing subpopulations with linear first order kinetics (Supplementary Fig. 17) leads to the following equations for the size x_i of subpopulation *i*:

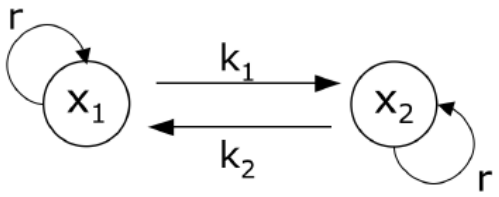
$$\begin{aligned} \dot{x}_1 &= r x_1 - k_1 x_1 + k_2 x_2 \\ \dot{x}_2 &= r x_2 + k_1 x_1 - k_2 x_2 \end{aligned} \tag{3}$$

where the dot denotes differentiation with respect to time, k_1 is the transition rate from x_1 to x_2 and vice versa for k_2. Since the cells are in a culture where there is a steady supply



Supplementary Figure 16. Time evolution of the weights for the two subpopulations as inferred by the GMM binning algorithm from the data. Symbols represent the weights $w_i(t)$ for G_1 (circles) and G_2 (triangles). The linear model is shown as a dotted line and the quadratic model as a solid line. The standard error for the weights obtained using Algorithm 1 is on the order of 10^{-3} and error bars are not shown. The dash-dotted black lines denote the stationary values of the weights (Supplementary Fig. 14). See also the caption of Fig. 2 in the main text for further discussion.

of nutrients, we assume that both subpopulations grow at the same rate r. This assumption is supported by the data in Supplementary Fig. 3, which indicates that cells grow at the same rate regardless of their Sca-1 levels.



Supplementary Figure 17. Linear model for two interacting and growing populations. The two subpopulations interact and cells transition from x_1 to x_2 at rate k_1 and vice versa at rate k_2.

The fluorescence intensity value is proportional to the relative fractions of the subpopulations. Therefore Equation (3) must be rewritten in terms of the relative populations. The evolution of the total population $y = x_1 + x_2$ is giving by $\dot{y} = ry$.
Let $w_1 = x_1 / y$ and $w_2 = x_2 / y$. From Equation (3), the evolution of w_i can be written as

$$\dot{w}_1 = \frac{\dot{x}_1 y - x_1 \dot{y}}{y^2} = -k_1 w_1 + k_2 w_2$$

whence we obtain

$$\dot{w}_1 = k_2 - (k_1 + k_2) w_1 \tag{4}$$

The solution for this equation is

$$w_1(t) = \frac{k_2}{k_1 + k_2} + e^{-(k_1 + k_2)t} \left(w_1(0) - \frac{k_2}{k_1 + k_2} \right) \tag{5}$$

The rates k_1 and k_2 and the integration constant $w_1(0)$ can be fitted to the data. The obtained fits are shown in Supplementary Table 2.

Note that the linear model does not capture two important features of the data. First, the asymptotic behavior of $w_1(\infty)$ and $w_2(\infty)$ is clearly different from the stationary *w* values in Supplementary Fig. 14. Second, Supplementary Fig. 16 shows that the linear model fails to capture the sigmoidal character of the growth for the earlier time points ($t \leq 96$) in the cases of the Sca-1^{Mid} and Sca-1^{High} population fractions.

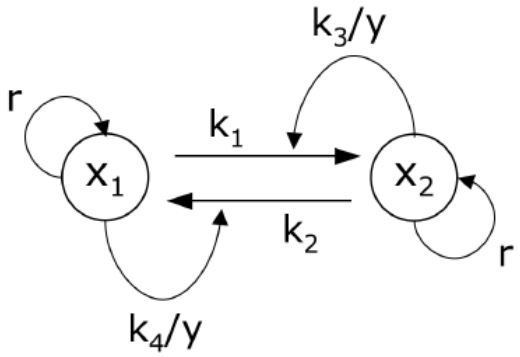
	Sca-1^{Low}	**Sca-1^{Mid}**	**Sca-1^{High}**
$k_1\left[h^{-1}\right]$	.0010	.0027	.0009
$k_2\left[h^{-1}\right]$	.0006	.0011	.0007
$w_1(0)$	.17	.0001	.0001
$w_1(\infty)$	.38	.30	.43
$w_2(\infty)$	.62	.70	.57

Supplementary Table 2. Parameters for the linear model for w_1 given by Equation (5)**.** The first three lines show the fitted parameters from the data. The last two lines show the asymptotic values calculated from the model.

(b) Nonlinear model

To better capture the asymptotic behavior and to explain the sigmoidal increase of the $w(t)$ for the Sca-1^{Mid} and Sca-1^{High} fractions, a simple non-linear model was introduced. The linear model in Equation (3) predicts an exponential behavior of the weights, compatible with a probabilistic (first-order) transition of individual cells from subpopulations G_2 to G_1. The sigmoidal departure from this exponential time evolution suggests a deviation from first order kinetics. In the simplest case, this can be caused by interaction between the cells. Cell differentiation and other discrete phenotypic state switches are often controlled by autocrine mechanisms that establish an autocatalytic loop that influences the rate of the state transition[11-13]. If, for instance, cells in one of the two states secrete a factor that promotes the switch to that state, this would cause the switching rate to depend on the ratio of the two subpopulations, resulting in sigmoidal rather than exponential kinetics.

As illustrated in Supplementary Fig. 18, the simplest model that captures this non cell-autonomous process contains two additional nonlinear (quadratic) terms that represent second order interactions between the two subpopulations. These terms model the effect of switching between the subpopulations mediated by the diffusion of a soluble signaling molecule. Assuming rapid diffusion, a simplified mean field model can be obtained in which the switching rate is proportional to the number of cells in a given state (subpopulation) in the culture.



Supplementary Figure 18. Nonlinear model of two interacting and growing populations. As in Supplementary Fig. 17, the two subpopulations interact and cells transition from G_1 to G_2 at rate k_1 and vice versa at rate k_2. In addition, the transition rates are increased by the terms representing the diffusive signaling interactions $k_3 x_2 / y = k_3 w_2$ and $k_4 x_1 / y = k_4 w_1$, where $y = x_1 + x_2$ is the total population.

The equations governing the growth of the two subpopulations will then be:

$$\begin{aligned} \dot{x}_1 &= \left[r x_1 - k_1 x_1 + k_2 x_2 \right] - k_3 w_2 x_1 + k_4 w_1 x_2 \\ \dot{x}_2 &= \left[r x_2 + k_1 x_1 - k_2 x_2 \right] + k_3 w_2 x_1 - k_4 w_1 x_2 \end{aligned} \tag{6}$$

where k_3 and k_4 are parameters determining the signal-induced switching rate. Rewriting in terms of the relative fraction w_1 as before, the growth is determined by

$$\dot{w}_1 = k_2 + (k - k_1 - k_2) w_1 - k w_1^2 \tag{7}$$

where $k = k_4 - k_3$. This equation has the solution

$$w_1(t) = \frac{1}{2} - \frac{k_1 + k_2}{2k} - \frac{\kappa}{2k} \tan\left(\frac{\kappa}{2} t - c_0 \right) \tag{8}$$

where $\kappa = \sqrt{k(2k_1 - 2k_2 - k) - (k_1 + k_2)^2}$ and $c_0 = \arctan\left(\left[k_1 + k_2 - k + 2k w_1(0) \right] / \kappa \right)$. The parameters obtained from fitting the data are shown in Supplementary Table 3.

Note that for the Sca-1$^{\text{Mid}}$ and Sca-1$^{\text{High}}$ experiments, $k_2 \ll k_1$. If we make $k_2 = 0$, Equation (7) becomes a standard logistic equation[14] with solution

$$w_1(t) = \frac{a w_1(0)}{k w_1(0) + (k - a w_1(0)) e^{-at}} \tag{9}$$

where $a = k - k_1$.

As shown in Supplementary Table 3 and Supplementary Fig. 16, the nonlinear model has the expected asymptotic behavior. Moreover, it captures the sigmoidal kinetics for the Sca-1^{Mid} and Sca-1^{High} datasets, with plateaus for both the early and late time points, with the correct asymptotic behavior.

	Sca-1^{Low}	**Sca-1^{Mid}**	**Sca-1^{High}**
$k_1\left[h^{-1}\right]$	.12	.09	.06
$k_2\left[h^{-1}\right]$	.012	1×10^{-8}	.0002
$k\left[h^{-1}\right]$	.084	.11	.08
$w_1(0)$	.17	.005	2×10^{-6}
$w_1(\infty)$	.20	.23	.21
$w_2(\infty)$	.80	.77	.79

Supplementary Table 3. Parameters for the quadratic model (7) **for** w_1**.** The first four rows show parameters fitted from the data. The last two rows show the asymptotic values calculated from the model with the fitted parameters.

We compare the linear and quadratic models to assess if the improvement of the fit warrants the introduction of the additional parameter in the quadratic model using the following formula of the AIC[15]:

$$AIC = N_d \log(MSE) + \frac{2n_p N_d}{N_d - n_p - 1} \tag{10}$$

where N_d is the number of data points used in the fitting, n_p is the number of independent parameters fitted, and MSE is the mean square error. Equation (10) is a standard form of AIC which approximates the likelihood in terms of the MSE assuming that the errors are normally distributed with a constant variance and corrects for the bias introduced when the number of data points is not much greater than the number of parameters[15].

For the Sca-1^{Mid} and Sca-1^{High} experiments, the AIC is smaller for the quadratic model, as shown in Supplementary Table 4. It is not surprising that there is no clear improvement of AIC for the Sca-1^{Low} experiment with the current amount and quality of data, since the data show a quasi-exponential decay. However, note that the asymptotic behavior of the quadratic model is more consistent with the stationary values of the experiments.

	Sca-1^{Low}	**Sca-1^{Mid}**	**Sca-1^{High}**
Linear	-70.2	-82.8	-88.5
Non-linear	-69.6	-86.0	-93.0

Supplementary Table 4. AIC for the linear and quadratic models. AIC values for the linear (5) and quadratic (7) models obtained from Equation (10) using the MSE calculated for the fitted functions with respect to the data. The values of the AIC indicate that for the Sca-1^{Mid} and Sca-1^{High} experiments, the improvement of the fit is large enough to warrant the introduction of the additional parameter in the quadratic model. Although the decay observed in the Sca-1^{Low} experiment is quasi-exponential, and can thus be fitted well by both models, the asymptotic value of the quadratic model is more consistent with the stationary data (Supplementary Tables 2 and 3). Note that comparing AIC values is only meaningful for related models and the absolute value of AIC does not carry any meaning. It is normal for AIC values to be either positive (Supplementary Fig. 13) or negative[15], as in this table.

Clearly, other nonlinear terms governing the switching could be considered to explain the features. However, due to our limited knowledge of the detailed genetic circuitry involved in this process, more elaborate models would be highly speculative. Equation (6) has the virtue of modeling a typical cell interaction (autocrine regulation) and making minimal assumptions about the nature of the interactions, while its parameters carry a distinct biological interpretation.

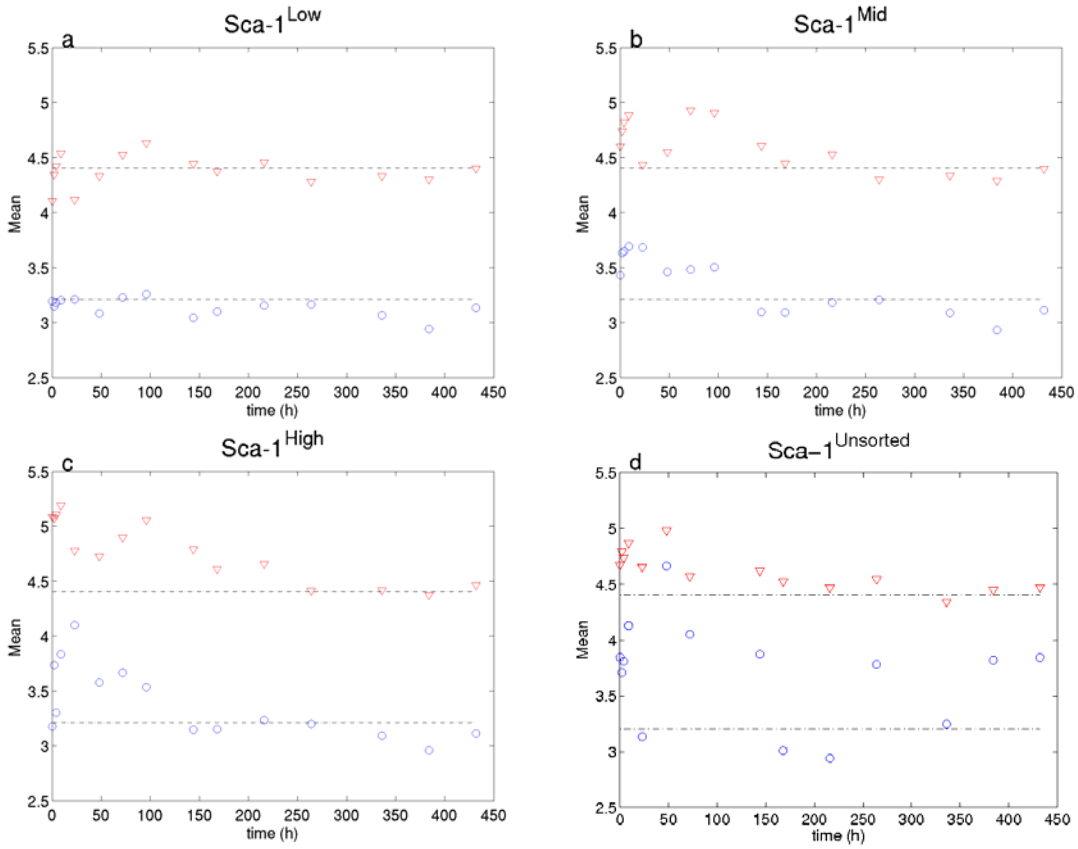
(c) Fast relaxation within subpopulations

The application of the GMM binning algorithm to the data provides us with an empirical decomposition into two (virtual) subpopulations for all times. Thus, it is possible to obtain the time evolution of the mean, variance and higher moments of these empirical sub-histograms.

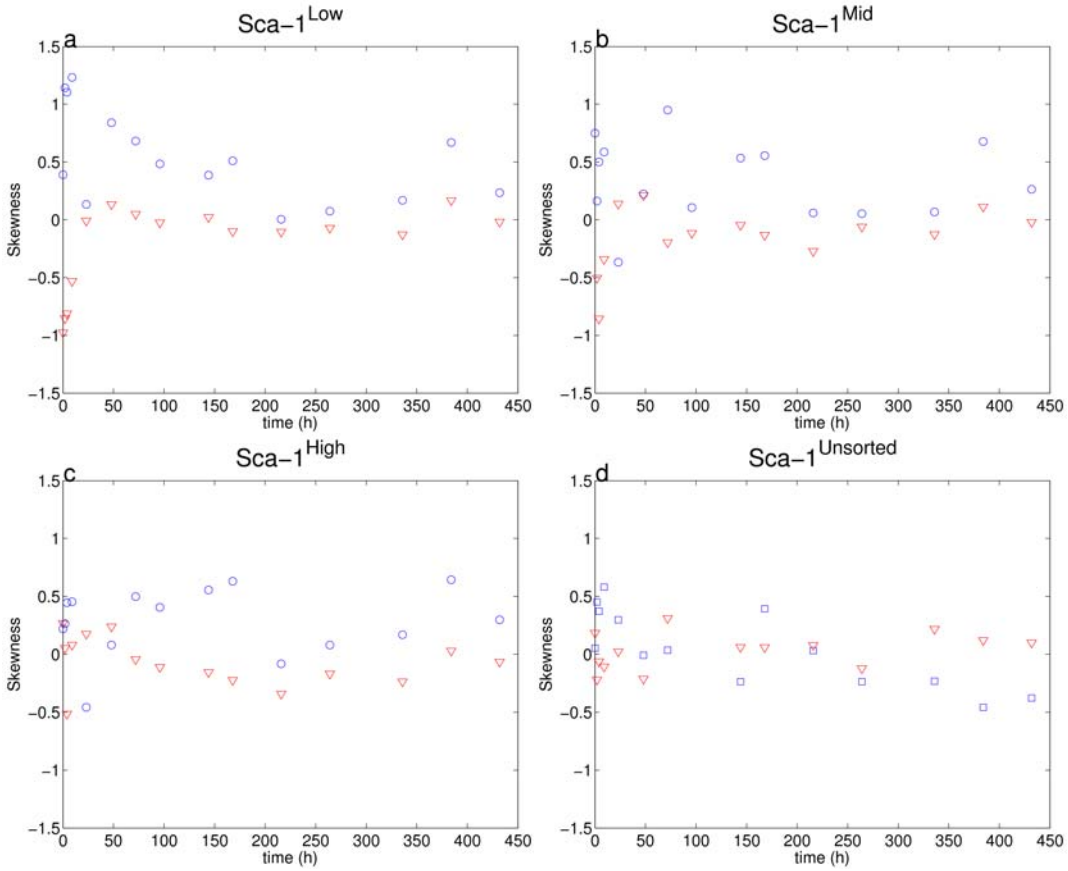
Supplementary Fig. 19 shows the time evolution of the means of the two Gaussians, μ_i, while Supplementary Fig. 20 shows the time evolution of the skewness of the distributions. For all three experiments, the cells spread out and repopulate the full width of G_2 over the first 24-48 hours after the sorting, as shown by the rapid disappearance of the skewness in Supplementary Fig. 20. The means of the Sca-1^{Mid} and Sca-1^{High} experiments exhibit decay towards the stationary value with a relaxation occurring on the order of around two days. This drift in mean fluorescence is due to the approach towards the stationary distribution from initial histograms that are not fully populated (Supplementary Fig. 15). The fluctuations in the means of the unsorted Sca-1 controls (Supplementary Fig. 19d) reveal an overall decay pattern reflecting procedural noise. No such pattern is visible in the evolution of the skewness which indicates that the problem is due to an overall shift of the histogram between different time-points.

The salient feature of the fluorescence histograms is that the changes of the relative heights of the peaks (the relative sizes of the two subpopulations) are much more significant than the changes of the width and locations of the two peaks. This is also in

agreement with the variability in the parameters of our GMM fits. As shown in Supplementary Fig. 19, the process of relaxation within the subpopulations is much faster than the slow process of balancing the weights to reconstitute the original distribution. This is in agreement with a model in which the restoration of the parental population distribution involves a more complex process with at least one discrete state transition rather than a simple mean-reverting process.



Supplementary Figure 19. Time evolution of the means for the two subpopulations as inferred by the GMM binning algorithm. Symbols represent the means $\mu_i(t)$ for G_1 (circle) and G_2 (triangle). The trend of the two populations is similar across all three experiments and it is due to the repopulation of the partly filled histograms. For the Sca-1^{Mid} and Sca-1^{High} experiments, there are few samples in G_1 for the early time-points and consequently, there are larger fluctuations. Comparison with the unsorted Sca-1 controls reveals a similar pattern of fluctuations which implies that they are due to procedural errors. The dash-dotted black lines denote the stationary values of the means (Supplementary Fig. 14).



Supplementary Figure 20. Time evolution of the skewness for the two subpopulations as inferred by the GMM binning algorithm. Symbols represent the skewness for G_1 (circle) and G_2 (triangle). The skewness, (i.e., the normalized third central moment of the distribution) measures the asymmetry of the distribution. A value of zero indicates perfect symmetry. After FACS sorting, the starting subpopulations are non-Gaussian but they rapidly become symmetric. The process of balancing the relative weights to reconstitute the parental population occurs at a much slower timescale, as shown in Supplementary Fig. 16. For the Sca-1^{Mid} and Sca-1^{High} experiments, there are few samples in G_1 for the early time-points and consequently, there are larger fluctuations.

S6. Supplementary Notes

1. Keyomarsi, K. et al. Synchronization of tumor and normal cells from G1 to multiple cell cycles by lovastatin. *Cancer Res* **51**, 3602-3609 (1991).
2. Zenger, V. E. et al. Quantitative flow cytometry: inter-laboratory variation. *Cytometry* **33**, 138-145 (1998).
3. Di Talia, S. et al. The effects of molecular noise and size control on variability in the budding yeast cell cycle. *Nature* **448**, 947-951 (2007).
4. Whitfield, M. L. et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* **13**, 1977-2000 (2002).
5. Pedraza, J. M. and van Oudenaarden, A. Noise propagation in gene networks. *Science* **307**, 1965-1969 (2005).
6. Mantzaris, N. V. From single-cell genetic architecture to cell population dynamics: quantitatively decomposing the effects of different population heterogeneity sources for a genetic network with positive feedback architecture. *Biophys J* **92**, 4271-4288 (2007).
7. Moon, T. K. The expectation-maximization algorithm. *Signal processing magazine, IEEE* **13**, 47-60 (1996).
8. Figueiredo, M. A. T. and Jain, A. K. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, 381-396 (2002).
9. Akaike, H. A new look at the statistical model identification. *IEEE Transactions on automatic control* **19**, 716-723 (1974).
10. Wedel, M. and Desarbo, W. S. A Mixture Likelihood Approach for Generalized Linear-Models. *Journal of Classification* **12**, 21-55 (1995).
11. Chitu, V. and Stanley, E. R. Colony-stimulating factor-1 in immunity and inflammation. *Curr Opin Immunol* **18**, 39-48 (2006).
12. Ruscetti, F. W., Akel, S., and Bartelmez, S. H. Autocrine transforming growth factor-beta regulation of hematopoiesis: many outcomes that depend on the context. *Oncogene* **24**, 5751-5763 (2005).
13. Whyatt, D. et al. An intrinsic but cell-nonautonomous defect in GATA-1-overexpressing mouse erythroid cells. *Nature* **406**, 519-524 (2000).
14. Verhulst, P.F. Notice sur la loi que la population suit dans son accroissement. *Curr. Math. Phys.* **10**, 113 (1838).
15. Burnham, Kenneth P. and Anderson, David R., *Model Selection and Multimodel Inference: A Practical Information - Theoretical Approach*. (Springer, 2002).