LETTERS

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[1,5] & Sui Huang[1]†

Phenotypic cell-to-cell variability within clonal populations may be a manifestation of 'gene expression noise'[1–6], or it may reflect stable phenotypic variants[7]. Such 'non-genetic cell individuality'[7] can arise from the slow fluctuations of protein levels[8] in mammalian cells. These fluctuations produce persistent cell individuality, thereby rendering a clonal population heterogeneous. However, it remains unknown whether this heterogeneity may account for the stochasticity of cell fate decisions in stem cells. Here we show that in clonal populations of mouse haematopoietic progenitor cells, spontaneous 'outlier' cells with either extremely high or low expression levels of the stem cell marker Sca-1 (also known as Ly6a; ref. 9) reconstitute the parental distribution of Sca-1 but do so only after more than one week. This slow relaxation is described by a gaussian mixture model that incorporates noise-driven transitions between discrete subpopulations, suggesting hidden multi-stability within one cell type. Despite clonality, the Sca-1 outliers had distinct transcriptomes. Although their unique gene expression profiles eventually reverted to that of the median cells, revealing an attractor state, they lasted long enough to confer a greatly different proclivity for choosing either the erythroid or the myeloid lineage. Preference in lineage choice was associated with increased expression of lineage-specific transcription factors, such as a >200-fold increase in Gata1 (ref. 10) among the erythroid-prone cells, or a >15-fold increased PU.1 (Sfpi1) (ref. 11) expression among myeloid-prone cells. Thus, clonal heterogeneity of gene expression level is not due to independent noise in the expression of individual genes, but reflects metastable states of a slowly fluctuating transcriptome that is distinct in individual cells and may govern the reversible, stochastic priming of multipotent progenitor cells in cell fate decision.

Cell-to-cell variability can be quantified by analysing the dispersion of expression levels of a phenotypic marker within a cell population. Flow cytometric analysis of EML cells, a multipotent mouse haematopoietic cell line[12], revealed an approximately 1,000-fold range in the level of the constitutively expressed stem-cell-surface marker Sca-1 among individual cells within one newly derived clonal cell population (Fig. 1a). The heterogeneity of Sca-1 expression in this clonal population was highly consistent between measurements (Fig. 1c) and could not be attributed to measurement noise (Fig. 1b). Moreover, cell-cycle-dependent cell size variation contributed only 1% to the observed variability of Sca-1 levels per cell (Supplementary Discussion and Supplementary Fig. 1).

To characterize the dynamics by which population heterogeneity arises, cells with the highest, middle and lowest ~15% Sca-1 expression level (denoted henceforth as Sca-1^{low}, Sca-1^{mid} and Sca-1^{high} fractions) were isolated from one clonal population using fluorescence-activated cell sorting (FACS). Cells were stripped free of the staining antibody immediately after isolation and were cultured in standard growth medium. Within hours, all three fractions showed broadening of the narrow Sca-1 histograms obtained immediately after sorting (Fig. 2a), but more than 9 days elapsed before the three fractions regenerated Sca-1 histograms similar to that of the parental (unsorted) population (Fig. 2a). Therefore, the restoration of the wide range of Sca-1 surface-expression levels is a slow process (requiring more than 12 cell doublings) that is independent of initial Sca-1 expression levels. Clonal heterogeneity was also regenerated from subclones derived from randomly selected individual cells that had varying initial mean Sca-1 levels (Supplementary Fig. 2).

What drives the regeneration of the parental 'bell-shaped' histogram from the three sorted population fractions (Fig. 2a)? Although a variety of mechanisms may in principle underlie this behaviour (Supplementary Discussion and Supplementary Fig. 3 and 4), we consider here a general theoretical stochastic formulation. Because the genetic circuitry governing the expression of Sca-1 is poorly understood[13], modelling the process explicitly with genetic circuits subjected to stochastic dynamics[14] is not feasible. Instead, we took a phenomenological approach to determine which general

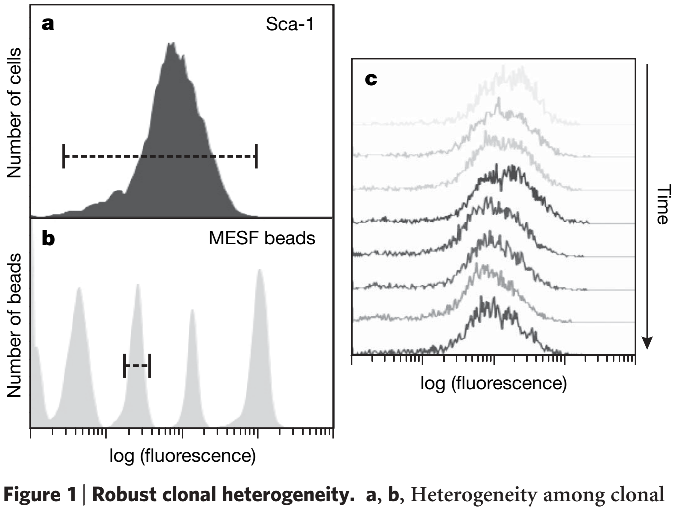

Figure 1 | Robust clonal heterogeneity. a, **b**, Heterogeneity among clonal cells in Sca-1 protein expression, detected by immunofluorescence flow cytometry (**a**), was significantly larger than the resolution limit of flow cytometry approximated by measurement of reference fluorescent MESF[24] beads (**b**). The dashed lines show the difference in spread of the distributions as explained in the text. **c**, Stability of clonal heterogeneity in Sca-1 over three weeks.

[1]Vascular Biology Programme, Department of Pathology and Surgery, Children's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Programme in Biophysics, [3]MD-PhD Programme, 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, UK. [5]Harvard Institute for Biologically Inspired Engineering, Cambridge, Massachusetts 02139, USA. †Present addresses: Department of Ophthalmology, Children's Hospital Boston, Boston, Massachusetts 02215, USA (M.H.); Institute for Biocomplexity and Informatics, University of Calgary, Calgary, Alberta T2N 1N4, Canada (S.H.).

class of models of stochastic processes best describes the observed behaviour. The simplest model would be an elementary mean-reverting (Ornstein–Uhlenbeck) process[15] that includes both noise-driven diffusion (capturing the generation of cell–cell variability) and a drift towards the deterministic equilibrium (representing relaxation to the parental distribution mean; Supplementary Theoretical Methods). However, a simple Ornstein–Uhlenbeck process describes the data only poorly, because it fails to recapitulate accurately the growth of the long left tail (for example, 100-fold range for the Sca-1$^{\text{high}}$ fraction) in the histogram.

An alternative explanation is that the relaxation process is complicated by slow dynamics on a rugged potential landscape that consists of multiple quasi-discrete state transitions, the stochastic nature of which produces an additional source of variability[16]. Recent analysis of human myeloid progenitor cells has provided experimental evidence for the existence of multiple metastable states[17], consistent with the dynamics of complex gene regulatory networks that control mammalian cell fates. We thus extended the simple Ornstein–Uhlenbeck model to include transitions between distinct states (virtual subpopulations) using a gaussian mixture model (GMM) as a first approximation to a multimodal system. As quantified by the Akaike information criterion (Supplementary Theoretical Methods), the data can be described by a minimal GMM model comprised of only two distinct states, each described as a gaussian, the parameters of which were obtained from the observed histograms in the stationary phase (time ≥ 9 days).

Our GMM model allowed us to partition cells in every measured histogram (time point) into two 'virtual subpopulations' (blue, subpopulation 1; red, subpopulation 2 in Fig. 2a) on the basis of the expression values of the individual cells, thus providing the time evolution of the mean μ_i and the relative abundance (weight) w_i for each subpopulation $i = 1, 2$ (Fig. 2b, c and Supplementary Theoretical Methods). This theoretical description suggests that the asymmetric broadening of the truncated histograms, as partially reflected in the changes in μ for the two subpopulations (Fig. 2b), only accounts for a fraction of the restoration of the equilibrium heterogeneity. In contrast, stochastic transitions between the subpopulations, as reflected by the evolution of the weights w_i, had a dominant role in the later relaxation to equilibrium. Importantly, for the Sca-1$^{\text{mid}}$ and Sca-1$^{\text{high}}$ fractions, changes in w_i were initially negligible until 96 h, at which point the w_i exhibited a steep change before eventually reaching a plateau (Fig. 2c).

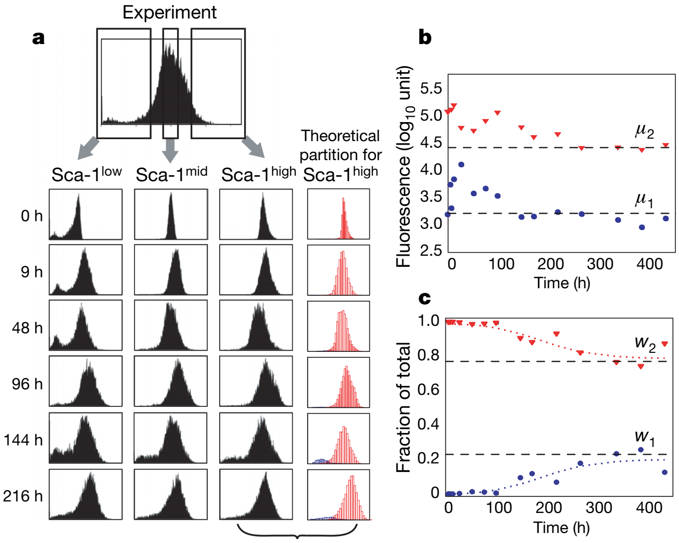

Figure 2 | Restoration of heterogeneity from sorted cell fractions. a, Clonal cells with the highest (Sca-1$^{\text{high}}$), middle (Sca-1$^{\text{mid}}$) and lowest (Sca-1$^{\text{low}}$) 15% Sca-1 expression independently re-established the parental extent of clonal heterogeneity after 216 h in separate culture. As an example, each cell in the Sca-1$^{\text{high}}$ experiment was theoretically partitioned into one of two GMM subpopulations (blue and red, right). **b**, **c**, The temporal evolution of the means $\mu_{1,2}$ (**b**) and weights $w_{1,2}$ (**c**) for the Sca-1$^{\text{high}}$ GMM subpopulations 1 and 2. The evolution of the weights was fitted to a sigmoidal function (**c**, dotted curves). Black dashed lines, equilibrium values for μ_i and w_i.

In summary, our results indicate that the observed clonal population heterogeneity of protein expression is not simply the manifestation of noise around a single, deterministic equilibrium (attractor) state described by an Ornstein–Uhlenbeck model. Instead, it is probably the result of processes involving stochastic state transitions in a system exhibiting multiple stable states[17], which may explain the slow regeneration of the parental heterogeneity.

These results suggest that whole-population averaging of the level of Sca-1 may not appropriately characterize its biological function. Instead, owing to the slowness of relaxation to the mean values, momentary levels of Sca-1 within individual cells may reflect distinct, enduring functional states that have different biological consequences. Thus, we asked whether clonal heterogeneity in Sca-1 protein expression correlates with heterogeneity of the differentiation potential of these cells. Indeed, among the secondary clones generated from the parental population, the rate of commitment to pro-erythrocytes in response to erythropoietin (Methods and Supplementary Fig. 5) was inversely correlated to the baseline mean Sca-1 expression of each clone (Supplementary Fig. 6). Similarly, for the three sorted fractions (Fig. 3a), the relative erythroid differentiation rates were distinct, with Sca-1$^{\text{low}}$ cells differentiating the fastest, followed by Sca-1$^{\text{mid}}$ and Sca-1$^{\text{high}}$ (Fig. 3b). Importantly, although the Sca-1$^{\text{low}}$ fraction differentiated into the erythroid lineage at a rate sevenfold higher than the Sca-1$^{\text{high}}$ fraction (Fig. 3b), the Sca-1$^{\text{low}}$ fraction was not composed of spontaneously and irreversibly precommitted pro-erythrocytes. Instead, these cells were still undifferentiated, as evidenced by expression of the stem cell marker c-kit (also known as Kit), their normal proliferation capacity (Supplementary Fig. 7) and their ability to reconstitute the parental histogram (Fig. 2a).

When we stimulated erythroid differentiation at various later time points after sorting, namely, on days 7, 14 and 21 of culture after sorting (as the Sca-1 histograms became more similar to each other while restoring the parental distribution), the difference in the erythroid differentiation rate between the Sca-1$^{\text{low}}$ and Sca-1$^{\text{high}}$ fractions was gradually lost (Fig. 3b–e). Surprisingly, despite the near complete convergence of the Sca-1 histograms at day 7, variability in differentiation kinetics was consistently detectable beyond 14 days after sorting (Fig. 3d). This suggests that clonal heterogeneity in Sca-1 expression controls differentiation potential but constitutes only a one-dimensional projection of separate states in the high-dimensional space of gene expression levels[17]. To reveal additional dimensions, we looked for correlated heterogeneity in other proteins and investigated whether expression of the erythroid-fate-determining transcription factor Gata1 (ref. 10) differed among the Sca-1 fractions. Real-time PCR revealed significantly higher *Gata1* messenger RNA levels in the erythroid differentiation-prone Sca-1$^{\text{low}}$ progenitor cells (260-fold increase over the Sca-1$^{\text{high}}$ fraction), followed by the Sca-1$^{\text{mid}}$ (2.7-fold increase over Sca-1$^{\text{high}}$ fraction) and Sca-1$^{\text{high}}$ fractions (Fig. 3g); these differences were paralleled by Gata1 protein levels (Fig. 3i). Importantly, *Gata1* mRNA expression among the three sorted fractions at 5 and 14 days after sorting (Supplementary Fig. 8) mirrored the gradual loss of variability observed in the differentiation kinetics for the erythroid lineage (Fig. 3b–e).

Gata1 has an antagonistic role to the myeloid-fate-determining transcription factor PU.1 in lineage determination; these two transcription factors mutually inhibit each other to regulate the erythroid versus myeloid fate decision[18]. Thus, we hypothesized that cells that are least prone to erythroid differentiation and exhibit low Gata1 expression may have high PU.1 levels, and thus be predisposed to the myeloid lineage. Indeed, real-time PCR revealed that Sca-1$^{\text{high}}$

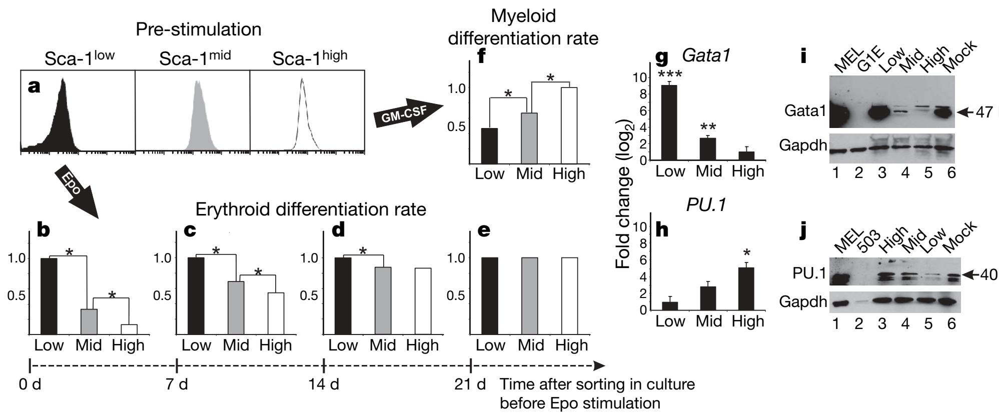

Figure 3 | Clonal heterogeneity governs differentiation potential. **a**–**f**, Sca-1^{low} (Low, black), Sca-1^{mid} (Mid, grey) and Sca-1^{high} (High, white) fractions (**a**) stimulated by erythropoietin (Epo, **b**) and GM-CSF (**f**) immediately after isolation showed variable differentiation rates into the erythroid and myeloid lineages, respectively. After 7, 14 and 21 days (d) of post-sort culture, erythropoietin-treated cells showed convergence in both pre-stimulation, baseline Sca-1 expression (Fig. 2a) and relative differentiation rates (**b**–**e**). Asterisk, $P < 0.001$ (two-tailed normal-theory test). **g**, **h**, Quantitative real-time PCR with reverse transcription analysis of *Gata1* (**g**) and *PU.1* (**h**) mRNA levels in Sca-1-sorted fractions. Means ± s.e.m. of triplicates shown. Triple asterisk, $P < 10^{-5}$; double asterisk, $P < 0.0002$; asterisk, $P < 0.003$ (one-tailed Student's *t*-test). **i**, **j**, Western blot analysis of Gata1 (**i**) and PU.1 (**j**) protein levels in Sca-1 fractions (lanes 3–5) and mock-sorted cells (lane 6). The MEL cell line (lane 1) was used as a positive control. G1E and 503 (lane 2) cell lines were negative controls for Gata1 and PU.1, respectively. Gapdh was the loading control.

progenitor cells have the highest *PU.1* mRNA levels (17-fold increase over Sca-1^{low} fraction), followed by the Sca-1^{mid} (3.6-fold increase over Sca-1^{low} fraction) and Sca-1^{low} fractions (Fig. 3h). These differences were paralleled by PU.1 protein levels (Fig. 3j). Furthermore, myeloid differentiation rate was the highest among Sca-1^{high} cells, followed by Sca-1^{mid} and Sca-1^{low} (Fig. 3f), in response to granulocyte–macrophage colony-stimulating factor (GM-CSF) and interleukin 3 (IL-3; Methods and Supplementary Fig. 5). These results show that within a clonal population of multipotent progenitor cells, spontaneous non-genetic population heterogeneity primes the cells for different lineage choices.

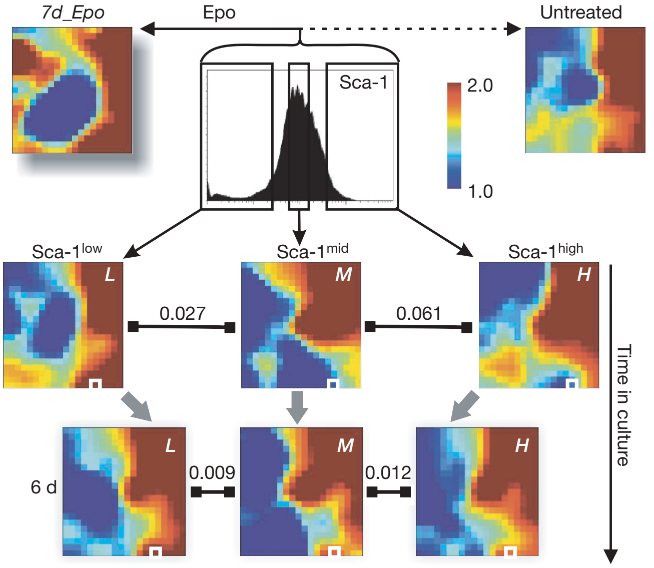

Figure 4 | Clonal heterogeneity of Sca-1 expression reflects transcriptome-wide noise. Self-organizing maps of global gene expression for a subset of 2,997 genes visualized with the GEDI[23] program for Sca-1^{low} (*L*), Sca-1^{mid} (*M*), Sca-1^{high} (*H*) fractions at 0 and 6 d after FACS isolation and for a differentiated erythroid culture (7 d erythropoietin, Epo) and an untreated control sample. Pixels in the same location within each GEDI map contain the same minicluster of genes. The colour of pixels indicates the centroid value of gene expression level for each minicluster in $\log_{10}$ units of signal. Dissimilarity between transcriptomes is indicated above the horizontal distance symbols. The Gata1-containing pixel is boxed in white.

Because both Gata1 and PU.1 are pivotal lineage-specific transcription factors, we asked whether the marked upregulation of Gata1 and associated downregulation of PU.1 in the most erythroid-prone Sca-1^{low} cells reflect a particular cellular state in terms of genome-wide gene expression. Microarray-based mRNA expression profiling on Sca-1^{low} (*L*), Sca-1^{mid} (*M*) and Sca-1^{high} (*H*) fractions immediately after sorting revealed that these three fractions differed considerably in their transcriptomes (Fig. 4). Replicate microarray measurements showed that the observed transcriptome differences could not be attributed solely to experimental error (Supplementary Fig. 9). Significance analysis of microarrays (SAM)[19] revealed >3,900 genes that were differentially expressed between the Sca-1^{low} and Sca-1^{high} fractions at a stringent false detection rate of 1.5%. The distinct global gene expression profiles of the three fractions converged to a common pattern within 6 days after sorting, a progression that can be quantified by the inter-sample distance metric $D = 1 - R$, where R is the Pearson correlation coefficient. The distances between the three profiles decreased from $D(L-M)_{0\text{ days}} = 0.027$ to $D(L-M)_{6\text{ days}} = 0.009$ and from $D(M-H)_{0\text{ days}} = 0.061$ to $D(M-H)_{6\text{ days}} = 0.012$ (Fig. 4 and Supplementary Table 1). Thus, the outlier populations reconstituted the traits of the parental population not only with respect to their distribution of Sca-1 expression (Fig. 2a) and differentiation rates (Fig. 3b–e) but also with respect to their gene expression profiles across thousands of genes. This global relaxation from both ends of the parental spectrum towards the centre is predicted by the model in which a stable cell phenotype, such as the progenitor state here, is a high-dimensional attractor state[20]. It also confirms that the Sca-1 outlier cells were not already irreversibly committed. Nevertheless, Sca-1^{low} cells exhibited a transcriptome that was clearly more similar than the Sca-1^{high} cells to the unsorted but maximally differentiated cells, achieved by culture in erythropoietin for 7 days (*7d_Epo*) (Fig. 4): $D(L - 7d\_Epo) = 0.079$ versus $D(H - 7d\_Epo) = 0.158$; Supplementary Table 1. This is a remarkable feat given the spontaneity and stochasticity of the process that generated these differentiation-prone outlier cells. In fact, with respect to 200 'differentiation marker genes' (Methods), only the Sca-1^{low} cells were statistically similar to the erythropoietin-treated cells ($P < 3 \times 10^{-14}$, pairwise *t*-test), whereas the Sca-1^{mid} ($P > 0.8$) and Sca-1^{high} ($P > 0.6$) cells were not, further confirming the transcriptome similarity between the Sca-1^{low} and erythropoietin-treated cells, which may be related to their increased Gata1 levels.

Our results demonstrate the robust nature of cell-to-cell variability that underlies the heterogeneity of gene expression in a clonal population of mammalian progenitor cells. Although the source of the heterogeneity and the molecular mechanisms responsible for its slow restoration remain to be elucidated, our experiments and general theoretical considerations point to discrete transitions in a dynamical system exhibiting multistability as one source of this behaviour. Independent of the specific mechanism, we show that biological function in metazoan cells is not necessarily determined by the ensemble average of a nominally homogenous cell population, and that outliers in a heterogeneous cell population do not simply represent irrelevant, short-lived phenotypic states caused by random fluctuations in the expression of a single gene. Instead, the departure from the average state is characterized by slowly fluctuating transcriptome-wide noise that has significant biological functionality in the priming of cell fate commitment. This finding helps unite two old dualisms: between plasticity and heterogeneity in explaining multipotency[21,22], and between instructive and selective regulation in explaining cell fate decisions[18]. Exploiting the spontaneous and transient yet enduring cell individuality in differentiation potential resulting from clonal heterogeneity also could be of practical value in attempts to steer lineage choice in stem cells for therapeutic applications.

METHODS SUMMARY

Creation of single-cell-derived subclones. Single-cell-derived subclones of EML cells were generated in three weeks by methylcellulose-plating at low cell densities, isolation of resulting colonies by hand with microscopic guidance, and expansion in liquid culture.

Flow cytometry and bead calibration. Cell surface protein immunostaining and flow cytometry measurements were performed using standard methods. For cells that were recultured after FACS, the staining antibody was removed as previously reported[17]. Quantum PE molecules of equivalent soluble fluorochrome (MESF) beads (Bangs Laboratories) were used to correct for daily fluctuations in flow cytometer sensitivity.

Gene expression profiling with microarrays. The MouseWG-6v1.1 Illumina microbead chips were used to perform gene expression profiling on total RNA extracted from FACS-sorted, or unsorted, cell populations.

Data analysis. Flow cytometry data were analysed using the software package FlowJo 2.2.2. Theoretical modelling and filtering of microarray data were performed with custom software written in Matlab 7.2. Statistical significance analysis of the microarray data was performed with the SAM[19] algorithm and self-organizing maps generated with the gene expression dynamics inspector (GEDI) software[23].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 27 January; accepted 31 March 2008.

1. Blake, W. J. *et al.* Noise in eukaryotic gene expression. *Nature* **422,** 633–637 (2003).
2. Elowitz, M. B. *et al.* Stochastic gene expression in a single cell. *Science* **297,** 1183–1186 (2002).
3. Pedraza, J. M. & van Oudenaarden, A. Noise propagation in gene networks. *Science* **307,** 1965–1969 (2005).
4. Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science* **304,** 1811–1814 (2004).
5. Rosenfeld, N. *et al.* Gene regulation at the single-cell level. *Science* **307,** 1962–1965 (2005).
6. Kaern, M. *et al.* Stochasticity in gene expression: from theories to phenotypes. *Nature Rev. Genet.* **6,** 451–464 (2005).
7. Spudich, J. L. & Koshland, D. E. Jr. Non-genetic individuality: chance in the single cell. *Nature* **262,** 467–471 (1976).
8. Sigal, A. *et al.* Variability and memory of protein levels in human cells. *Nature* **444,** 643–646 (2006).
9. van de Rijn, M. *et al.* Mouse hematopoietic stem-cell antigen Sca-1 is a member of the Ly-6 antigen family. *Proc. Natl Acad. Sci. USA* **86,** 4634–4638 (1989).
10. Cantor, A. B., Katz, S. G. & Orkin, S. H. Distinct domains of the GATA-1 cofactor FOG-1 differentially influence erythroid versus megakaryocytic maturation. *Mol. Cell. Biol.* **22,** 4268–4279 (2002).
11. Koschmieder, S. *et al.* Role of transcription factors C/EBPα and PU.1 in normal hematopoiesis and leukemia. *Int. J. Hematol.* **81,** 368–377 (2005).
12. Tsai, S. *et al.* Lymphohematopoietic progenitors immortalized by a retroviral vector harboring a dominant-negative retinoic acid receptor can recapitulate lymphoid, myeloid, and erythroid development. *Genes Dev.* **8,** 2831–2841 (1994).
13. Holmes, C. & Stanford, W. L. Concise review: stem cell antigen-1: expression, function, and enigma. *Stem Cells* **25,** 1339–1347 (2007).
14. Guido, N. J. *et al.* A bottom-up approach to gene regulation. *Nature* **439,** 856–860 (2006).
15. Uhlenbeck, G. E. & Ornstein, L. S. On the theory of Brownian Motion. *Phys. Rev.* **36,** 823–841 (1930).
16. Kurchan, J. & Laloux, L. Phase space geometry and slow dynamics. *J. Phys. Math. Gen.* **29,** 1929–1948 (1996).
17. Chang, H. H. *et al.* Multistable and multistep dynamics in neutrophil differentiation. *BMC Cell Biol.* **7,** 11 (2006).
18. Huang, S. *et al.* Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev. Biol.* **305,** 695–713 (2007).
19. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98,** 5116–5121 (2001).
20. Huang, S. *et al.* Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett.* **94,** 128701 (2005).
21. Enver, T., Heyworth, C. M. & Dexter, T. M. Do stem cells play dice? *Blood* **92,** 348–351; discussion 352 (1998).
22. Orkin, S. H. & Zon, L. I. Hematopoiesis and stem cells: plasticity versus developmental heterogeneity. *Nature Immunol.* **3,** 323–328 (2002).
23. Eichler, G. S., Huang, S. & Ingber, D. E. Gene expression dynamics inspector (GEDI): for integrative analysis of expression profiles. *Bioinformatics* **19,** 2321–2322 (2003).
24. Zenger, V. E. *et al.* Quantitative flow cytometry: inter-laboratory variation. *Cytometry* **33,** 138–145 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was funded by grants to S.H. from the Air Force Office of Scientific Research and, in part, from the National Institutes of Health. H.H.C. is partially supported by the Presidential Scholarship and the Ashford Fellowship of Harvard University. M.H. and M.B. are supported by the Life Sciences Interface and Mathematics panels of the Engineering and Physical Sciences Research Council of the UK. D.E.I. is supported by the National Health Institutes and the Army Research Office. We thank K. Orford, P. Zhang, A. Mammoto, J. Daley, J. Pendse and M. Shakya for experimental assistance, and W. Press and K. Farh for discussions.

Author Contributions H.H.C. designed the study, performed the experiments, analysed the data, participated in the theoretical analysis and drafted the manuscript. M.H. constructed the theoretical model and performed the theoretical analysis. M.B. constructed the model, supervised the work and revised the manuscript. D.E.I. supervised the work and revised the manuscript. S.H. conceived of the study, designed experiments, supervised the work, participated in the experimental and theoretical analysis and drafted the manuscript. All authors read and approved the final manuscript.

Author Information The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) under the GEO Series accession number GSE10772. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.H. (sui.huang@ucalgary.ca).

METHODS

Culture of EML cells, derivation of subclones, and differentiation. EML cells[26] (a gift from K. Orford and D. Scadden) were maintained in growth medium containing Iscove's modified Dulbecco's medium (IMDM), 20% horse serum, 12–15% (v/v) medium conditioned (CM) by baby hamster kidney (BHK) cells producing murine kit-ligand (MKL), and 1% glutamine/penicillin/streptomycin. To obtain single-cell-derived subclones, cells were plated into 60-mm plates at 500–2,000 cells ml^{-1} density in 1% methylcellulose (Methocult M3134) containing growth medium and incubated without disturbance for 10 days. Individual well-demarcated colonies were hand-picked with Pasteur pipettes under microscopic guidance and were transferred to liquid cultures in microwell plates. Typical subclones required ~18 days in culture to expand to a sufficiently large population for the experiment. To differentiate EML cells into the erythroid lineage, a previously reported differentiation protocol[12] was adapted. In brief, on day 1, cells were cultured in growth medium plus 10 ng ml^{-1} mouse recombinant erythropoietin (Sigma-Aldrich) at 250,000 cells ml^{-1} density. On day 3, cells were spun down and re-suspended into IMDM plus 20% horse serum, 2% BHK/MKL-CM and 10 ng ml^{-1} mouse recombinant erythropoietin at 125,000 cells ml^{-1} density to give resulting erythroid cells a growth advantage. One day 6, an additional 10 ng ml^{-1} of erythropoietin was added. Typically, 7 days of erythropoietin treatment generated ~40–60% (of total) pro-erythrocytes that were benzidine-stain-positive and Sca-1/c-kit double-negative (Supplementary Fig. 5). Benzidine staining was performed following a reported protocol[25] and examined by microscopy after cytospin. To differentiate EML cells into myeloid cells, a previously reported differentiation protocol[12] was adapted. In brief, on day 1, cells were cultured in growth medium plus 10 ng ml^{-1} mouse recombinant IL-3 (Peprotech) and 10^{-5} M retinoic acid (Sigma-Aldrich) at 300,000 cells ml^{-1} density. On day 4, cells were washed thoroughly with PBS to remove remaining SCF from the growth medium and were cultured in IMDM plus 20% horse serum, 2% BHK/MKL-CM, 10 ng ml^{-1} mouse recombinant GM-CSF (R&D Systems) and 10^{-5} M retinoic acid (Sigma-Aldrich) at 200,000 cells ml^{-1} density. On day 6, an additional 10 ng ml^{-1} GM-CSF was added. After 7–9 days, differentiated myeloid cells dominate the culture and show Mac-1 (Itgam, integrin α M) and Gr-1 (Ly6G) expression by flow cytometry.

Flow cytometry, FACS and bead calibration. For direct cell-surface-protein immunostaining, the antibodies Sca-1–PE (Caltag) and c-kit–FITC (BD Pharmingen) were used at 1:1,000 dilutions in ice-cold PBS plus 1% fetal calf serum with (flow cytometry) or without (FACS) 0.01% $\mathrm{NaN_3}$. Appropriate isotype control antibodies (BD Pharmingen) were used to establish background signal caused by non-specific antibody binding. Propidium iodide staining was correlated with lower forward scatter among EML cells (Supplementary Fig. 10). Thus, dead cells with positive propidium iodide staining were easily removed from all analysis by gating out the low forward scatter population. Flow cytometry was performed on a Becton Dickinson FACSCaliber analyser and FACS with either a Becton Dickinson FACSAria or an AriaSpecial Sorter ultraviolet laser system at the Dana Farber Cancer Institute Flow Cytometry Core. Computational data analysis was done with FlowJo 2.2.2. For cell sorting, input cell number ranged from 60×10^6 cells to 100×10^6 cells. Cells were sorted into ice-cold medium for a maximal duration of 3 h. Gates for the lowest, middle and highest Sca-1 expressors were set based on the proportion of total population. For cells that were re-cultured after FACS, the staining antibody was removed following a previously reported protocol[17]. Quantum PE MESF beads (Bangs Laboratories) were used to correct for the effect of day-to-day fluctuations in the flow cytometer following the manufacturer's instructions. Calibration curves were constructed using Matlab 7.2 (MathWorks) and were used to convert obtained fluorescence data into absolute MESF units for the purpose of quantitative theoretical modelling.

Gene expression profiling with microarrays and data analysis. Gene expression profiling was performed at the Molecular Genetics Core facility at the Children's Hospital Boston using MouseWG-6 v1.1 microbead chips (Illumina). Raw gene expression data were first subjected to rank-invariant normalization using BeadStudio 3.0. Matlab 7.2 was used to filter the list of 46,628 genes on the basis of two sets of criteria. First, detection P-value based on Illumina replicate gene probes: genes with detection P-values > 0.01 in all samples were called 'absent' in all samples and thus removed (giving rise to set 1, consisting of 14,038 genes). Genes with differing 'detection call' ('absent' versus 'present') between the duplicate samples were also removed. Second, fold-change: genes that did not show at least a twofold change compared to the Sca-1$^{\mathrm{mid}}$ fraction in 4 out of the 12 total samples were also removed (resulting in set 2: 2,997 genes). Alternatively, the SAM[19] algorithm was used to filter by fold change at a stringent false detection rate of 1.5% (resulting in set 3: 3,973 genes). Qualitative conclusions did not depend on the exact stringency of the filtering. After filtering, gene expression levels were transformed by $\log_{10}$ and subjected to clustering analyses. GEDI maps for visual representation of global gene expression based on self-organizing maps were generated using the program GEDI[23] (http://www.childrenshospital.org/research/ingber/GEDI/gedihome.htm). In GEDI, each 'tile' within a 'mosaic' represents a minicluster of genes that have highly similar expression pattern across all the analysed samples. The same genes are forced to the same mosaic position for all GEDI maps, hence allowing direct comparison of transcriptomes based on the overall mosaic pattern. The colour of tiles indicates the centroid value of gene expression level for each minicluster. Dissimilarity between samples was quantified by $1 - R$, where R is the Pearson's correlation coefficient calculated for all genes in a pair of samples. For statistical analysis of the similarity between the sorted fractions and the erythropoietin-treated sample, a subset of ~200 'differentiation marker genes' were obtained from stringent SAM analysis of the unsorted, untreated control and the unsorted, erythropoietin-treated sample.

25. Wang, R., Clark, R. & Bautch, V. L. Embryonic stem cell-derived cystic embryoid bodies form vascular channels: an *in vitro* model of blood vessel development. *Development* **114**, 303–316 (1992).

26. Tsai, S. *et al.* Lymphohematopoietic progenitors immortalized by a retroviral vector harboring a dominant-negative retinoic acid receptor can recapitulate lymphoid, myeloid, and erythroid development. *Genes Dev.* **8**, 2831–2841 (1994).