

# Decoding the regulatory network of early blood development from single-cell gene expression measurements

Victoria Moignard<sup>1,2,11</sup>, Steven Woodhouse<sup>1,2,11</sup>, Laleh Haghverdi<sup>3,4</sup>, Andrew J Lilly<sup>5</sup>, Yosuke Tanaka<sup>1,2,6</sup>, Adam C Wilkinson<sup>1,2</sup>, Florian Buettner<sup>3</sup>, Iain C Macaulay<sup>7</sup>, Wajid Jawaid<sup>1</sup>, Evangelia Diamanti<sup>1,2</sup>, Shin-Ichi Nishikawa<sup>6</sup>, Nir Piterman<sup>8</sup>, Valerie Kouskoff<sup>5</sup>, Fabian J Theis<sup>3,4</sup>, Jasmin Fisher<sup>9,10</sup> & Berthold Göttgens<sup>1,2</sup>

Reconstruction of the molecular pathways controlling organ development has been hampered by a lack of methods to resolve embryonic progenitor cells. Here we describe a strategy to address this problem that combines gene expression profiling of large numbers of single cells with data analysis based on diffusion maps for dimensionality reduction and network synthesis from state transition graphs. Applying the approach to hematopoietic development in the mouse embryo, we map the progression of mesoderm toward blood using single-cell gene expression analysis of 3,934 cells with blood-forming potential captured at four time points between E7.0 and E8.5. Transitions between individual cellular states are then used as input to develop a single-cell network synthesis toolkit to generate a computationally executable transcriptional regulatory network model of blood development. Several model predictions concerning the roles of Sox and Hox factors are validated experimentally. Our results demonstrate that single-cell analysis of a developing organ coupled with computational approaches can reveal the transcriptional programs that underpin organogenesis.

Blood has long served as a model to study organ development owing to the accessibility of blood cells and the availability of markers for specific cell populations. Blood development initiates at gastrulation from multipotent Flk1<sup>+</sup> (encoded by *Flk1*, also known as *Kdr*) mesodermal cells, which initially have the potential to form blood, endothelium and smooth muscle cells<sup>1,2</sup>. Blood development represents one of the earliest stages of organogenesis, as the production of primitive erythrocytes is required to support the growing embryo. Single-cell gene expression analysis has already been successfully applied to study the earliest stages of preimplantation mouse and human development<sup>3–5</sup>, to identify lineage commitment<sup>6</sup> and transcriptional regulatory<sup>7</sup> events in blood, and, more recently, to probe the emergence of hematopoietic stem cells (HSCs) from the hemogenic endothelium of the dorsal aorta<sup>8</sup>.

Here we report *in vivo* gene expression analysis of early blood development at the single-cell level, focusing on transcription factors as regulators of cell fate. Using qRT-PCR, we analyzed >40 genes in 3,934 cells with blood and endothelial potential from five populations at four sequential stages of post-implantation mouse development between embryonic day (E)7.0 and E8.25. We adapted the diffusion plot methodology previously reported in nonbiological contexts<sup>9</sup> for

dimensional reduction of single-cell data, where pseudotemporal ordering of individual cells revealed a putative developmental hierarchy branching toward both blood and endothelial-like fates. To discover the underlying regulatory network, we developed a single-cell network synthesis (SCNS) toolkit for synthesizing executable Boolean network models from binary single-cell expression states, which correspond to the on and off patterns of transcription factor expression. Using this toolkit we identified a core network of 20 highly connected transcription factors, which could reach eight stable states representing blood and endothelium. We validated model predictions to demonstrate that Sox7 blocks primitive erythroid development, and Sox and Hox factors directly regulate expression of the HSC regulator, *Erg*. The SCNS toolkit therefore opens up network reconstruction for other systems without the requirement for prior knowledge of regulatory interactions.

## RESULTS

### Capturing cells with blood potential during gastrulation

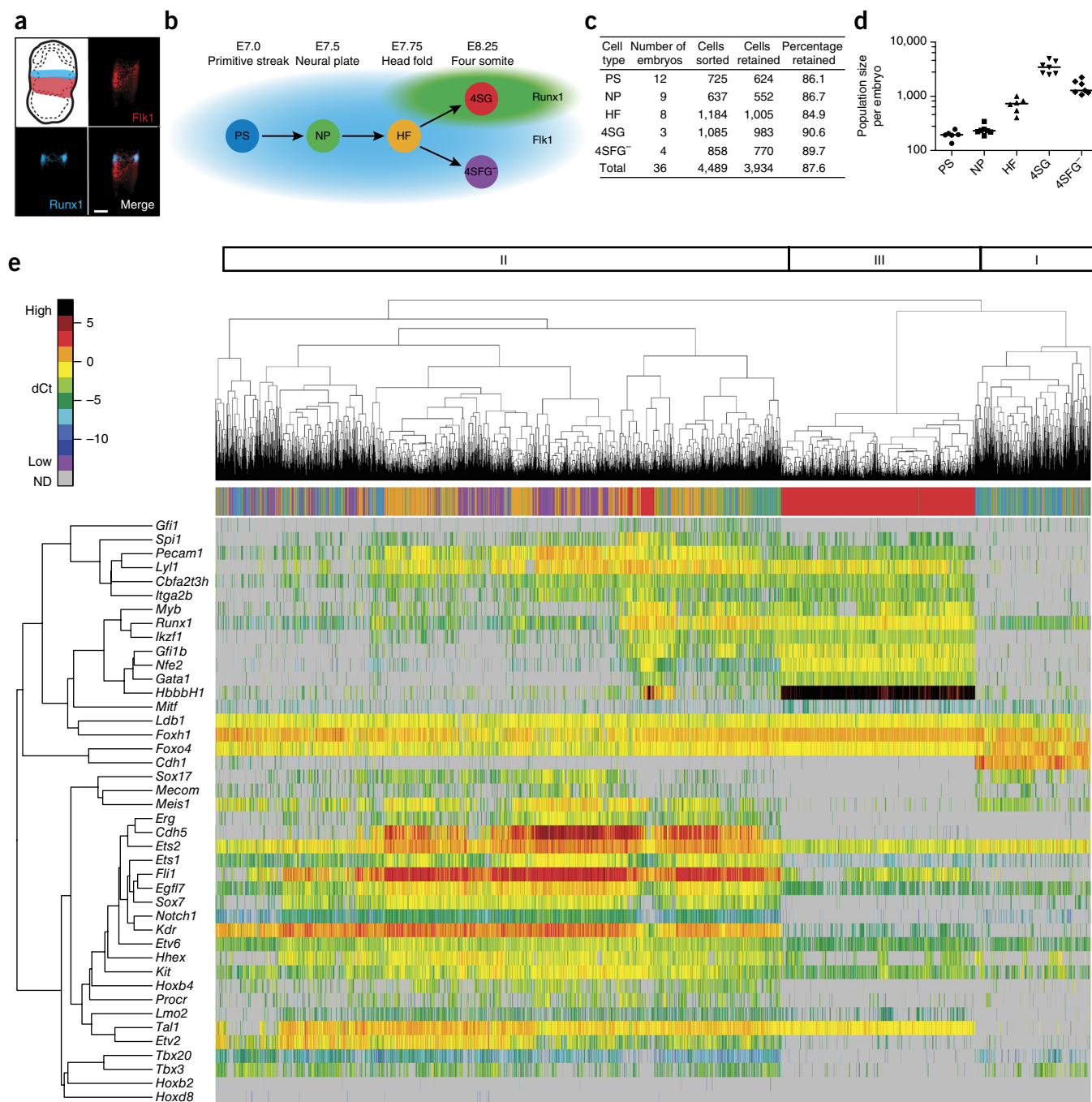
The first wave of primitive hematopoiesis originates from Flk1<sup>+</sup> mesoderm<sup>1,2,10</sup>, with all hematopoietic potential in the mouse contained within the Flk1<sup>+</sup> population from E7.0 onwards. Although some blood

<sup>1</sup>Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK. <sup>2</sup>Wellcome Trust - Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. <sup>3</sup>Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany. <sup>4</sup>Department of Mathematics, Technische Universität München, Garching, Germany. <sup>5</sup>Cancer Research UK Stem Cell Haematopoiesis Group, Paterson Institute for Cancer Research, University of Manchester, Manchester, UK. <sup>6</sup>Laboratory for Stem Cell Biology, RIKEN Center for Developmental Biology, Chuo-ku, Kobe, Japan. <sup>7</sup>Sanger Institute-EBI Single Cell Genomics Centre, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. <sup>8</sup>Department of Computer Science, University of Leicester, Leicester, UK. <sup>9</sup>Microsoft Research Cambridge, Cambridge, UK. <sup>10</sup>Department of Biochemistry, University of Cambridge, Cambridge, UK. <sup>11</sup>These authors contributed equally to this work. Correspondence should be addressed to B.G. (bg200@cam.ac.uk) or J.F. (jasmin.fisher@microsoft.com or jf416@cam.ac.uk).

Received 15 August 2014; accepted 16 January 2015; published online 9 February 2015; doi:10.1038/nbt.3154

progenitor cells lose Flk1 expression just before the onset of circulation<sup>11</sup>, previous work using a *LacZ* reporter knocked into the *Runx1* locus showed that hematopoietic potential remains confined to the *Runx1*<sup>+</sup> fraction<sup>12</sup>, which was confirmed with a GFP reporter driven by the *Runx1* +23 enhancer, which reproduces *Runx1* expression<sup>8</sup>. Using Flk1 expression in combination with a *Runx1*-ires-GFP reporter mouse<sup>13</sup> therefore allowed us to capture cells with blood potential at

distinct anatomical stages across a time course of mouse development (Fig. 1a,b). Single Flk1<sup>+</sup> cells were flow sorted at E7.0 (primitive streak, PS), E7.5 (neural plate, NP) and E7.75 (head fold, HF) stages. We subdivided E8.25 cells into putative blood and endothelial populations by isolating GFP<sup>+</sup> cells (four somite, 4SG) and Flk1<sup>+</sup>GFP<sup>-</sup> cells (4SFG<sup>-</sup>), respectively (Fig. 1b and Supplementary Fig. 1a). Cells were sorted from multiple embryos at each time point, with



**Figure 1** Single-cell gene expression analysis of early blood development. (a) Flk1 and Runx1 staining in E7.5 mesoderm and blood band, respectively. Scale bar, 100  $\mu$ m. (b) Single cells sorted from five populations at four anatomically distinct stages from E7.0–E8.25. (c) Quantification of cells sorted and retained for analysis after quality control. (d) Quantification of Flk1<sup>+</sup>, GFP<sup>+</sup> or Flk1<sup>+</sup>GFP<sup>-</sup> cells in embryos at each time point from FACS data (Supplementary Fig. 1a). Line indicates median. (e) Unsupervised hierarchical clustering was performed using the Spearman correlation and complete linkage for the normalized gene expression of the 33 transcription factors and 7 marker genes in all cells. Shown is the level of expression for each gene in every cell (see key). Rows represent genes and columns represent cells, with the top colored bar indicating the embryonic stage of origin for each cell (see b for color code). Major cell clusters I, II and III are indicated. ND, not detected.

**Figure 2** Diffusion plots identify developmental trajectories. Diffusion plot of all 3,934 cells calculated from the expression of 33 transcription factors and seven marker genes (top left). The expression levels of individual genes were then overlaid onto the diffusion plot to highlight patterns of expression (see **Supplementary Fig. 5** for additional genes). Circle, PS; diamond, NP; triangle, HF; cross, 4SG; square, 4SFG (visible in high resolution version of figure).

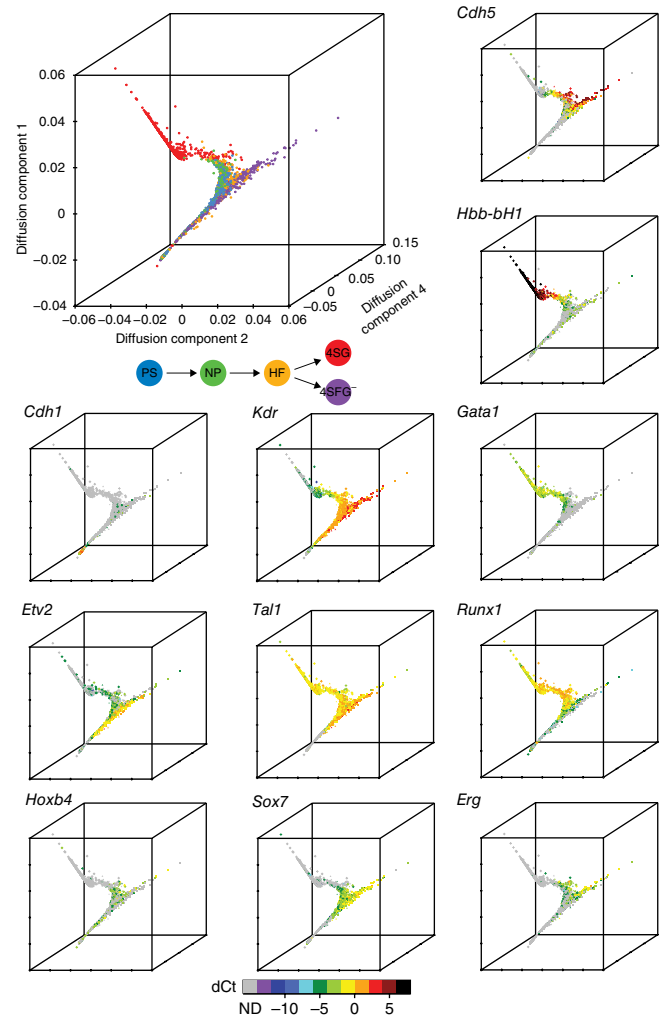
3,934 cells going on to subsequent analysis (**Fig. 1c**). Total cell numbers (**Supplementary Fig. 1b**) and numbers of cells of appropriate phenotypes (**Fig. 1d**) present in each embryo were estimated from fluorescence-activated cell sorting (FACS) data, indicating that for the first three stages, more than one embryo equivalent of  $\text{Flk1}^+$  cells was collected.

We next quantified the expression of 33 transcription factors involved in endothelial and hematopoietic development<sup>14</sup>, nine marker genes, including the embryonic globin *Hbb-bH1* and cell surface markers such as *Cdh5* (VE-Cadherin) and *Itga2b* (CD41), as well as four reference housekeeping genes in all 3,934 cells using microfluidic qRT-PCR technology<sup>7</sup> (**Supplementary Table 1**), which resulted in >150,000 quantitative expression scores.

### Development of blood progenitor cells is not synchronized

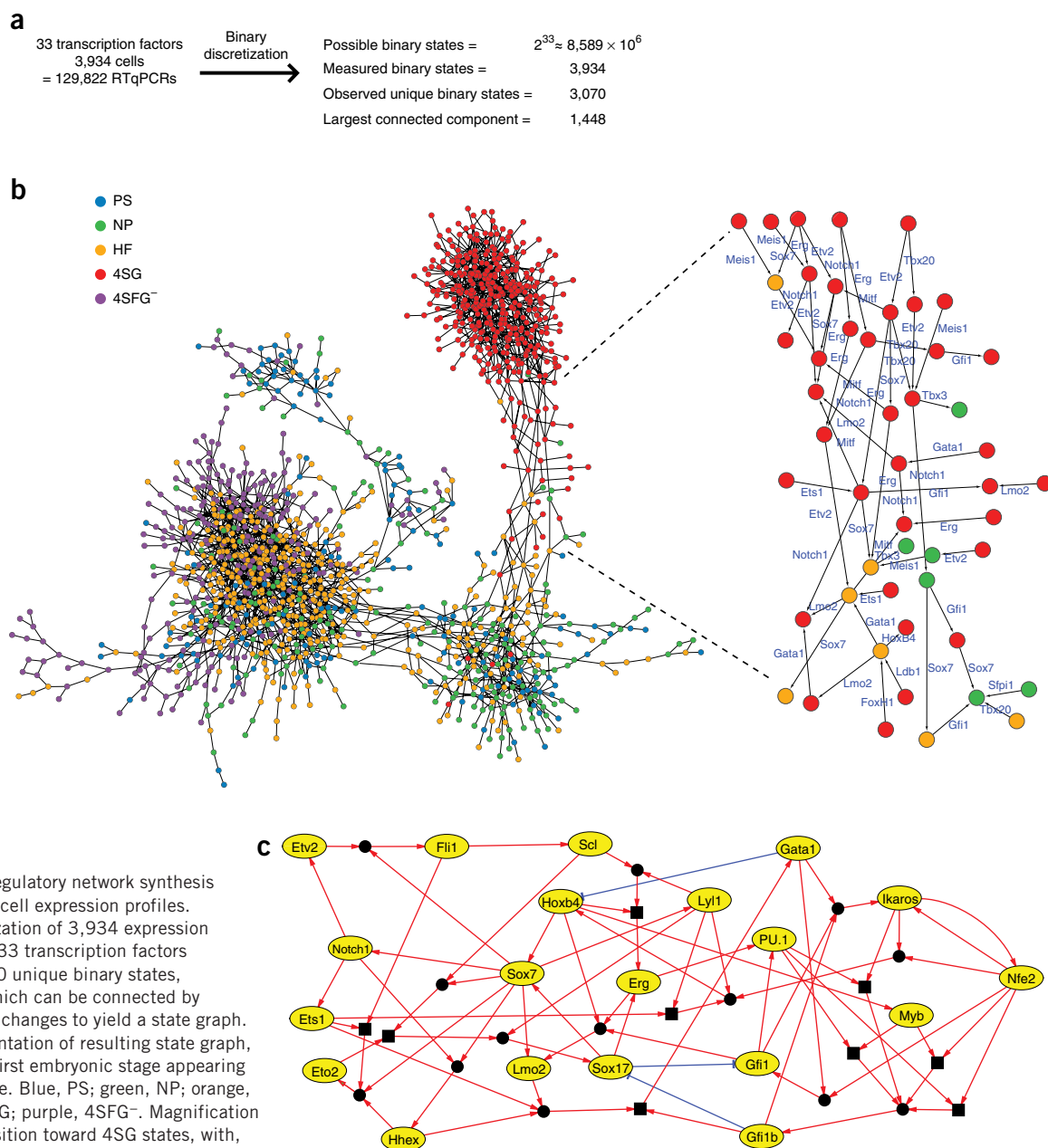
Unsupervised hierarchical clustering of the 33 transcription factor and 9 marker genes across all 3,934 cells revealed three major clusters (**Fig. 1e**). Cluster I was small and comprised mostly PS and NP cells. It did not express blood-associated genes, but showed low expression of some endothelial genes and high expression of *Cdh1* (E-cadherin), likely representing mesodermal cells at the primitive streak<sup>15</sup>. Cluster II contained the greatest number of cells and included most of the PS, NP, HF and 4SFG<sup>+</sup> cells, was characterized by endothelial gene expression, and contained subclusters with elevated expression of hemogenic endothelial genes, such as *Cdh5*, or hematopoietic genes such as *Gfi1*, indicating that this cluster contains a continuum of cells maturing from mesodermal to hematopoietic and endothelial fates. Cluster III was formed by most of the E8.25  $\text{Runx1}^{\text{GFP}+}$  4SG cells, and had robust expression of hematopoietic genes (including *Hbb-bH1*, *Gata1*, *Nfe2*, *Gfi1b*, *Ikzf1* (Ikaros) and *Myb*), and low expression of endothelial genes (*Erg*, *Sox7*, *Sox17*, *Hoxb4*, *Cdh5*). The mixing of cells from different anatomical stages by hierarchical clustering analysis therefore suggested that developmental maturation of single cells in early mesodermal cell populations is asynchronous, with cells at multiple stages expressing similar combinations of developmental regulators. This is consistent with the gradual ingression of cells through the primitive streak and lineage commitment during gastrulation.

Principal component analysis (PCA) of the expression values of all 3,934 cells confirmed the large-scale mixing of cells from different anatomical stages, with only 4SG cells forming a stage-specific group (**Supplementary Fig. 2a**). The PCA was retrospectively colored to show which embryo each cell belongs to (**Supplementary Fig. 2b**), to determine whether this mixing is the result of developmental asynchrony within embryos or differences in maturation between different embryos classified as being of the same anatomical stage. We quantified the percentage of cells from each embryo belonging to clusters I, II and III, as identified by hierarchical clustering (**Fig. 1e** and **Supplementary Fig. 2c,d**). This showed that cells collected from each embryo at the PS, NP and HF stages were distributed across clusters I and II, with the earlier stages showing a greater bias toward cluster I than later stages. These results are therefore consistent with a model whereby cells representing both early and later stages along the differentiation trajectory toward blood are present throughout



the PS, NP and HF time points, captured as snapshot measurements in our high-throughput, single-cell expression profiling.

A proportion of  $\text{Flk1}^+$  cells will give rise to mesodermal lineages other than blood and endothelium, and the extent to which they emerge over time and contribute to the variability would need to be analyzed using different gene sets. Notably, however, >50% of PS, NP and HF cells expressed both *Flk1* and *Runx1* at the mRNA level, highlighting the presence of  $\text{Flk1}^+$  cells with hemogenic potential<sup>8,12</sup> from the earliest time points (**Supplementary Fig. 3**). Analysis of 50-cell pools from the PS, NP and HF stages by RNA-seq showed graded expression increases of hematopoietic and endothelial genes from the E7.0 to the E7.5 and E7.75 samples. This is entirely consistent with the continuous emergence of blood-specified cells deduced from our single-cell data, as an increase in the proportion of cells expressing a given gene between stages will increase population-averaged expression measurements. Key mesodermal and cardiac genes, by contrast, showed graded downregulation in the pooled-cell RNA-seq (**Supplementary Fig. 4**). These graded expression changes over time are not consistent with a discrete on or off switch at a specific developmental time point, but could again be due to gradual changes in the proportion of cells expressing the marker genes, similar to our observations from single-cell analysis of blood and endothelial genes. Alternatively, quantitative changes in expression levels within a constant proportion of cardiac-specified cells would similarly result in a change in the overall expression level of a population and cannot



**Figure 3** Regulatory network synthesis from single-cell expression profiles.

(a) Discretization of 3,934 expression profiles for 33 transcription factors yields 3,070 unique binary states, 1,448 of which can be connected by single-gene changes to yield a state graph.

(b) Representation of resulting state graph, colored by first embryonic stage appearing in each state. Blue, PS; green, NP; orange, HF; red, 4SG; purple, 4SFG<sup>-</sup>. Magnification of fate transition toward 4SG states, with, for example, Sox7 expression switching off along all routes.

(c) Representation of synthesized asynchronous Boolean network models for core network of 20 transcription factors. Red edges indicate activation; blue edges indicate repression. Square boxes represent AND operations. Circles connecting edges indicate multiple update rules.

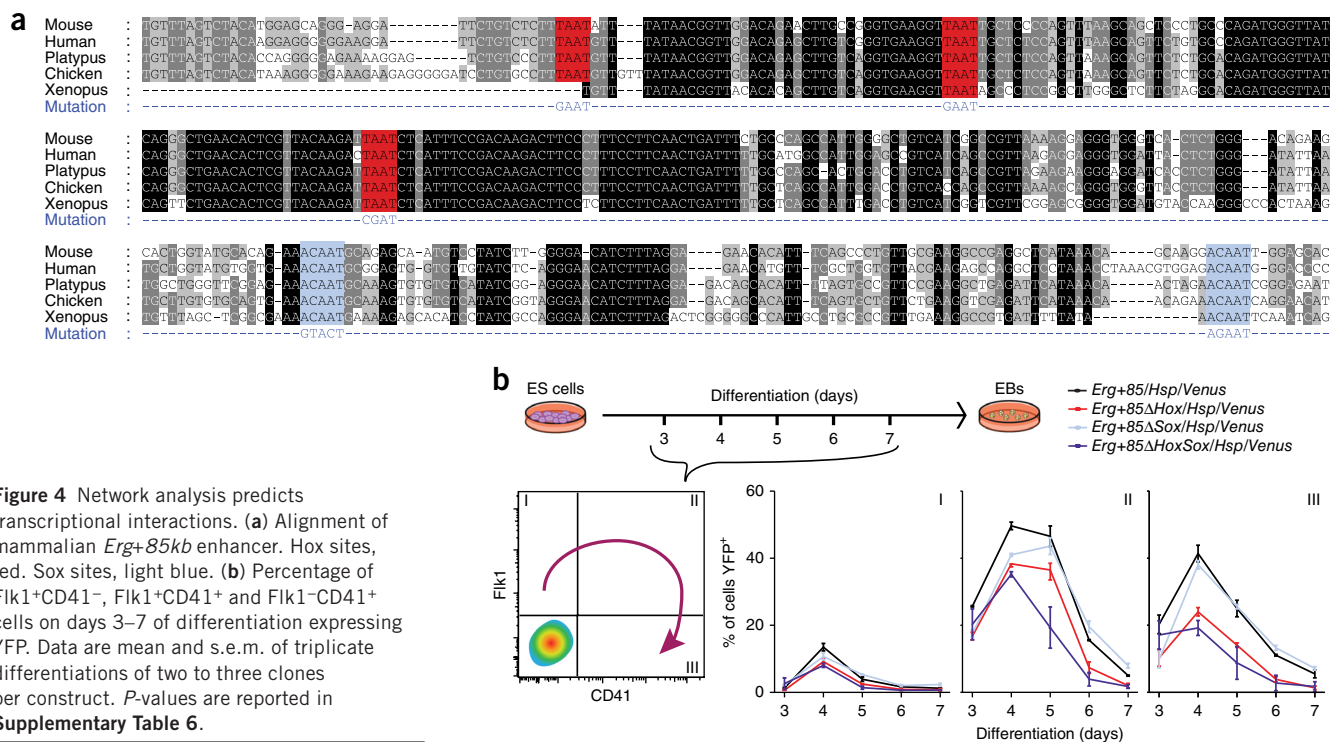
be excluded from the pooled-cell RNA-seq. Therefore, our results indicate, at least for cells destined to become blood and endothelium, that these cells arise at all stages of the analyzed time course rather than in a synchronized fashion at one precise time point, consistent with the gradual nature of gastrulation. Notably, only single-cell analysis over a developmental time-course has the power to reveal the contribution to cellular heterogeneity made by unsynchronized maturation of individual cells.

### Diffusion maps identify developmental trajectories

To identify and visualize putative developmental trajectories from the PS to 4S stages in the single-cell gene expression data, we developed a computational approach for dimension reduction. Our method is based on the concept of diffusion distances, which can be

interpreted as a metric for objects (here, cells) that are related to each other through a gradual but stochastic, diffusion-like process, such as cellular differentiation. In brief, similarities between all 3,934 cells are calculated based on their gene expression patterns, and then visualized globally in a three-dimensional (3D) map (Fig. 2 and Supplementary Fig. 5). The resulting components span a low-dimensional diffusion-space, in which distance reflects how similar cells are in terms of their diffusion distance, and can be inferred to represent developmental time.

Although there is extensive mixing between PS, NP, HF and 4SFG<sup>-</sup> populations in the diffusion plot, there is a general progression in the cell stages present in different regions of the plot from largely early E7.0 PS and E7.5 NP cells through the later HF cells to the E8.25 4SG cells that form a homogeneous cluster, in line with the expected



**Figure 4** Network analysis predicts transcriptional interactions. **(a)** Alignment of mammalian *Erg+85kb* enhancer. Hox sites, red. Sox sites, light blue. **(b)** Percentage of Flk1+CD41<sup>-</sup>, Flk1+CD41<sup>+</sup> and Flk1-CD41<sup>+</sup> cells on days 3–7 of differentiation expressing YFP. Data are mean and s.e.m. of triplicate differentiations of two to three clones per construct. *P*-values are reported in **Supplementary Table 6**.

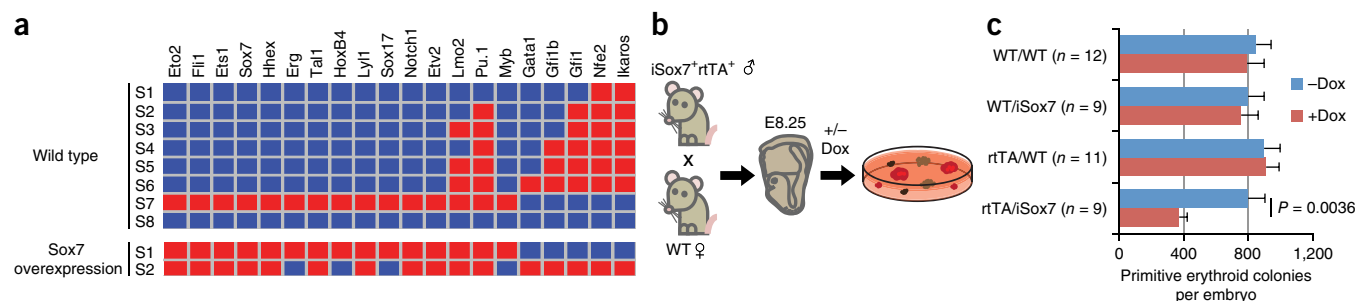
developmental progression of the blood system or 4SFG<sup>-</sup> cells (see **Supplementary Fig. 6** for projection of individual populations). Furthermore, we observed that whereas the E8.25 Flk1<sup>+</sup>Runx1-GFP<sup>-</sup> (4SFG<sup>-</sup>) cells mostly mix with earlier Flk1<sup>+</sup> cells, a subset that was not identified by clustering or PCA branches off. This branch expresses endothelial and hemogenic endothelial genes (*Cdh5*, *Erg*, *Itga2b*, *Pecam1* (CD31), *Sox7*, *Fli1*) with lower to absent expression of *Etv2* and *Runx1*. This observation is consistent with the known bifurcation of blood and endothelium (reviewed in ref. 16) and the downregulation of *Runx1* in more mature endothelial cells<sup>17</sup>. This bifurcation was more apparent in the diffusion maps than by PCA, independent component analysis or t-SNE (**Supplementary Fig. 7**). Genes that mark early, intermediate and late stages of blood development showed dynamic expression across the diffusion map (**Fig. 2**), with *Cdh1* expressed first, followed by *Cdh5* and then the embryonic globin *Hbb-bH1*. The transcription factors *Etv2*, *Tal1* (Scl), *Runx1* and *Gata1* were expressed in a pattern consistent with their known sequential roles during the development of hemangioblasts through to erythroid cells<sup>18–25</sup>. Dynamic expression patterns were also observed for other transcription factors not previously recognized as major regulators of primitive hematopoiesis, including *Erg*, *Sox7* and *Hoxb4*. The diffusion map method therefore represents an attractive approach for ordering cells in developmental time, identifying patterns of expression for key regulators and bifurcation events not readily found with standard algorithms.

### Synthesis of a network model for early blood development

The correspondence between the diffusion map and known developmental timelines suggested that the measured expression changes reflect developmental trajectories and might be exploited to define the regulatory networks that drive mesodermal cells toward a hematopoietic fate. Cell fate decisions have been modeled successfully using state space analysis of asynchronous Boolean regulatory network models<sup>26,27</sup>. In this approach, each gene is associated with a Boolean variable (1 or 0), which represents whether the gene is expressed or not

expressed, respectively, in the cell. Each gene is also given a Boolean update rule that specifies how its expression value changes over time owing to regulation by other genes. Boolean network dynamics are then modeled by a series of asynchronous single-gene changes, and state space analysis reveals the final stable states of the model. We were interested in the inverse problem: if we think of the single-cell expression profiles as the state space of a Boolean network, can we identify the underlying gene regulatory logic? Although single-cell data have been used to refine static networks curated from the literature<sup>28</sup>, to our knowledge Boolean rules have not been derived directly from single-cell expression data without a priori knowledge of the structure of the network. To tackle this complex question of revealing the molecular changes underpinning cell state transitions, we developed the SCNS toolkit to synthesize Boolean networks based on single-gene transitions in our data.

We first discretized all 3,934 single-cell expression profiles to binary states and connected those states that differ in the expression of only one gene. The threshold for binary discretization was determined as described in Online Methods. This yielded a connected state transition graph of 1,448 expression states, connected by single-gene transitions (**Fig. 3a,b**). The number of times each state occurs is indicated in **Supplementary Figure 8**. The probability of seeing even one repeated state or neighbor in the whole theoretical state space is negligible, illustrating the nonrandom nature of the data. Most states that corresponded to the Runx1-GFP<sup>+</sup> 4SG cells clustered together at one end of the state transition graph, whereas states corresponding to cells from other time points were dispersed between two additional clusters. Likely developmental transitions were revealed, with specific genes consistently switching on or off along all routes linking the major clusters. We therefore considered this state transition graph as a possible representation of developmental expression state changes based on single-gene switches, and next asked whether this could be used for regulatory network reconstruction. Notably, analysis of real and simulated populations of 20 cells showed that pools for the same stage clustered closely together, which masked variation and therefore



**Figure 5** *In silico* perturbations predict key regulators of blood development. (a) Network stable states for wild-type and Sox7 overexpression. Red indicates expressed; blue indicates not expressed. (b) Colony assays with or without doxycycline from genotyped E8.25 embryos from iSox7<sup>+</sup>rtTA<sup>+</sup> mice crossed with wild types. (c) Quantification of primitive erythroid colonies after 4 days (mean and s.e.m. for the number of embryos indicated). *P*-value was determined using the student's *t*-test for the number of embryos indicated.

would not have provided the number of transcriptional states required for network synthesis (Supplementary Fig. 9).

The direction of movement between two states in the state transition graph is initially not defined. Our method assigns a direction to each connection based on overall movement from the early PS to the later 4SG states, and then finds Boolean update functions for each gene that are consistent with its expression changes across the entire transition graph. Unlike previous analyses of single-cell gene expression data, which have largely relied on statistical properties of the data viewed as a whole, our method can recover mechanistic logic and determine the direction of interactions. When the method was applied to our data set, we obtained a core network of 20 transcription factors with endothelial and blood-associated gene modules centered on Sox7, Hoxb4 and Erg, and on Gata1 and PU.1 (also known as Spi1), respectively. For some genes, there were multiple possible consistent update functions. For example, there are two solutions for Erg, both of which include activation by Hoxb4 and Sox17. In total there were 39 possible functions, an average of two per gene. This led to 46,656 possible models from the different combinations of the 39 update rules (Fig. 3c and Supplementary Table 2). Repeating the network synthesis with bootstrapping and a different discretization threshold demonstrated the robustness of our protocol (Supplementary Tables 3 and 4).

### Network synthesis predicts direct regulation of Erg

We next asked whether links in our single-cell expression-derived network models can reveal direct regulatory interactions. To provide support for our model, we identified high-confidence gene regulatory regions in the gene loci of the 20 transcription factors in our network by interrogating a compendium of transcription factor ChIP-seq data from hematopoietic cell types<sup>29</sup>, followed by identification of binding sites for the 20 transcription factors within these regions (Supplementary Fig. 10). 27 of the 39 Boolean rules (70%) are supported by the presence of evolutionarily highly conserved motifs for the upstream regulators in the target gene locus (Supplementary Table 2), with support for at least one Boolean rule for 16/20 transcription factors. This finding suggested that many of the regulatory interactions proposed in our model may be direct upstream regulator/downstream target gene relationships. To provide further validation, we focused on Erg, which our models predicted is activated by Sox17, or by Hoxb4 in combination with Lyl1 or Scl (Tal1). By analyzing a Hoxb4 ChIP-seq data set<sup>30</sup>, we showed that Hoxb4 can bind to the Erg+85kb enhancer (Supplementary Fig. 11a), which we previously showed to be active in blood stem and progenitor cells<sup>31,32</sup>. Moreover, comparative sequence analysis revealed that the Erg+85kb contains highly conserved Hox and Sox binding sites (Fig. 4a).

To investigate regulation of Erg by Hox and Sox factors, we took advantage of a recently described embryonic stem cell-based reporter system in which single-copy enhancer transgenes linked to the Hsp68/Venus reporter are targeted to the Hprt locus<sup>33</sup>, allowing robust comparisons of wild-type and mutant enhancer activity during *in vitro* differentiation. We tracked enhancer activity during embryoid body differentiation, where cells transit from a Flk1<sup>+</sup>CD41<sup>-</sup> mesoderm/hemangioblast state, through a Flk1<sup>+</sup>CD41<sup>+</sup> intermediate, to a Flk1<sup>-</sup>CD41<sup>+</sup> hematopoietic state<sup>33–36</sup>. Flow cytometric analysis revealed a dynamic pattern of YFP expression for the wild-type enhancer, peaking at days 4–5 and highest in the Flk1<sup>+</sup>CD41<sup>+</sup> population (Fig. 4b and Supplementary Fig. 11b,c). Similar expression was seen in the Sox mutant, whereas mutation of the Hox motifs caused a reduction of YFP<sup>+</sup> cells, and the combined Hox and Sox mutant reduced the proportion of YFP<sup>+</sup> cells further still. We also saw similar patterns of expression in the other populations, which constitute a larger proportion of the embryoid body cells but have a lower percentage of YFP<sup>+</sup> cells (Fig. 4b and Supplementary Fig. 11b,c). This suggests that Hox and Sox factors activate and maintain Erg expression largely independently and additively. When abstracted to the Boolean level, this result is therefore more consistent with the OR logic in our network than with the alternative AND logic, where single mutations would result in an effect as strong as the combined mutant.

### Model execution reveals key switches during development

Next, we assessed whether our network models faithfully recapitulate blood and cardiovascular development, in which endothelial and primitive blood cells emerge from a common mesodermal progenitor. To do this, we determined the stable states of the network model that correspond to those expression patterns for the 20 transcription factors that satisfy all the Boolean network rules, and therefore can remain stable. We found that only eight stable states are reachable in total across all possible models, including “endothelial-like” (WT-S7) and “blood-like” expression states (WT-S2 to S6) (Fig. 5a). Of note, 432 models had both the endothelial-like state and at least one of the blood-like states (WT-6) as stable states, thus capturing the functionality of bipotential Flk1<sup>+</sup> precursors.

Finally, we explored the consequences of *in silico* perturbation. Overexpression and knockout experiments were simulated for each transcription factor and the ability of the network to reach wild-type or new stable states was assessed (Supplementary Table 5). For a number of factors, stable states 6 or 7 were no longer reachable. Among these, enforced expression of Sox7, a factor normally downregulated when cells transit toward the 4SG state (Fig. 3b), resulted in the stabilization of the endothelial module and an inability

to reach any of the blood-like states (Fig. 5a). Only two stable states were possible, among the lowest for all factors, and furthermore, Sox7 is predicted to regulate more targets than any other transcription factor, suggesting that perturbing its expression could have important downstream consequences (Supplementary Table 5). To validate this prediction, we crossed the previously reported iSox7<sup>+</sup>rtTA<sup>+</sup> male mice<sup>37</sup> with wild-type females, collected embryos at E8.25 and performed colony-forming assays (Fig. 5b). Embryos carrying both transgenes showed a 50% reduction of primitive erythroid colony formation and simultaneous appearance of undifferentiated hemangioblast-like colonies following doxycycline-induced Sox7 expression compared to controls (Fig. 5c and data not shown). This suggests, in agreement with modeling data and gene expression patterns, that downregulation of Sox7 is important for the specification of primitive erythroid cells.

## DISCUSSION

Determining the structure and function of transcriptional regulatory networks is crucial to advancing our understanding of developmental and disease processes and is therefore a key aim of stem and developmental biology. However, studies to date have mainly used population-based data for network construction or have focused on statistical properties of populations of single cells for network inference.

Bayesian network methods provide a very computationally efficient approach to inferring causal relationships among a set of variables and have previously been applied to infer cellular signaling networks from single-cell data<sup>38</sup>. However, these approaches infer a directed acyclic graph where there is no feedback between nodes, a limitation not shared by our approach. In addition, the inference of edges is reliant on network interventions in which many different cell populations are generated by experimentally perturbing genes, and the differences between these populations are used to infer causality. Generating such intervention data is very time consuming and cannot be done when studying wild-type *in vivo* development. Instead, researchers typically look at the pairwise correlation of genes across single-cell measurements<sup>7,39</sup>. For example, partial correlation analysis measures the degree of association between two genes while controlling for potential effects of all other genes<sup>40</sup>. We performed this analysis (Supplementary Fig. 12), and found agreement with many of the edges in our synthesized network; however, this analysis failed to predict the Sox/Hox regulation of *Erg*, which we validated experimentally. Moreover, connections do not specify which gene is the upstream regulator and which is the downstream target, and therefore do not reveal mechanistic logic.

To our knowledge no previous study has analyzed the development of an entire mammalian organ at single-cell resolution. Here we have studied the earliest stages of blood development from mesoderm through to the emergence of primitive erythroid cells, and demonstrate that single-cell expression profiling, coupled with computational approaches for network synthesis, can reveal molecular control mechanisms of mammalian organogenesis. Analysis of 46 genes in blood precursors across 1.25 days of post-implantation mouse embryonic development showed that cellular maturation may be asynchronous, with individual cells maturing at different speeds and a large proportion expressing both *Flk1* and *Runx1*, indicating that they are committing to hemogenic endothelial development. The graded changes in expression for key regulators of other mesodermal fates seen in the cell pools analyzed by RNA-seq are also consistent with cells expressing the gene emerging over the time-course analyzed, although alternative explanations such as changes in the level of expression cannot be excluded. Furthermore, our diffusion map

methodology highlighted the hierarchical nature of organ development, with waves of transcription factor and marker expression and a bifurcation at the four-somite stage. The presence of embryonic globin and erythroid transcription factor *Gata1* in one branch and endothelial markers such as *Pecam1* and *Cdh5* in the other suggests that this bifurcation represents the separation of blood and endothelial fates<sup>14,16</sup>. Trapnell *et al.*<sup>41</sup> recently reported an exciting method related to our diffusion map approach for the analysis of single-cell, RNA-seq, time-course data, where construction of a minimum spanning tree ordered differentiating cells in developmental pseudotime. Although the authors suggested that this methodology could be used to map regulatory networks, such results were not included in their paper. Moreover, cells were sampled from cells differentiating *in vitro* rather than directly from embryos.

Here we achieved reconstruction of regulatory network models by deriving expression-state graphs from high-throughput, single-cell, gene expression profiling data and using the expression-state graphs to determine gene regulatory rules. First, gene expression profiles are discretized to binary expression states, where 1 represents a gene that is expressed and 0 represents no measurable expression. Then, pairs of states are connected if they differ in the expression state of exactly one gene, resulting in a state graph. Finally, Boolean rules are found for each gene, which allow a walk from early states to late states by means of a series of single-gene transitions. The result is a set of Boolean rules matching the experimental data that can be combined into a network model. This method is provided as the SCNS toolkit. It requires no prior knowledge of regulatory interactions but instead derives its logic directly from the gene expression data.

We followed this method of network synthesis with steady state and *in silico* perturbation analyses that identified blood and endothelial-like expression patterns and implicated Sox7 in the regulation of erythroid fate, which we subsequently validated using transgenic mouse assays. Network synthesis also identified several previously known transcription factor interactions, including close linkage of *Etv2*, *Fli1* and *Tal1*, where the latter two are known to function downstream of *Etv2* in the hemangioblast<sup>42,43</sup>. To test whether our network model reveals additional direct interactions, we focused on *Erg*, an essential transcription factor for definitive hematopoiesis and adult HSC function<sup>44,45</sup>. Our network predicted that *Erg* expression can be activated either by Sox17 or Hoxb4. The *Erg*+85kb enhancer was previously shown to be controlled by Ets and Gata factors and to be active during hematopoietic development<sup>32</sup> and in HSCs<sup>31</sup>. However, neither Hox nor Sox transcription factors had been implicated in *Erg*+85kb activity.

Sox7 and Sox17 belong to the SoxF family of transcription factors, which have recently been shown to confer arterial identity in combination with RBPJ/Notch<sup>46</sup>. Arterial identity is linked with the blood-forming potential of hemogenic endothelial cells in the embryo. Moreover, Hoxb4 expression is also known to enhance blood potential<sup>47</sup>, yet there is very little knowledge about how SoxF factors or Hoxb4 integrates into the wider network regulating blood development. Our integrated approach of single-cell expression profiling coupled with network synthesis and subsequent experimental validation identifies *Erg* as a downstream target of Sox and Hox factors during early blood specification. Coupled with our observations here that downregulation of Sox7 is a key event in the development of primitive erythroid cells, our study demonstrates how network modeling from single cells can help to reveal the transcriptional hierarchies that control mammalian development. Rapid technological advances in our ability to perform single-cell profiling<sup>48,49</sup> suggest that this approach will be widely applicable to other organ systems and may also inform the development of improved cellular programming strategies.

## METHODS

Methods and any associated references are available in the [online version of the paper](#).

**Accession codes.** GEO: [GSE61470](#).

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

## ACKNOWLEDGMENTS

We thank J. Downing (St. Jude Children's Research Hospital, Memphis, TN, USA) for the Runx1-ires-GFP mouse. Research in the authors' laboratory is supported by the Medical Research Council, Biotechnology and Biological Sciences Research Council, Leukaemia and Lymphoma Research, the Leukemia and Lymphoma Society, Microsoft Research and core support grants by the Wellcome Trust to the Cambridge Institute for Medical Research and Wellcome Trust - MRC Cambridge Stem Cell Institute. V.M. is supported by a Medical Research Council Studentship and Centenary Award and S.W. by a Microsoft Research PhD Scholarship.

## AUTHOR CONTRIBUTIONS

V.M., A.J.L., Y.T., A.C.W. and I.C.M. performed experiments and analyzed data. S.W., L.H., F.B. and N.P. developed computational tools. W.J. and E.D. analyzed data. S.-I.N., V.K., F.J.T., J.F. and B.G. conceived the study. V.M., S.W., J.F. and B.G. wrote the paper with help from all co-authors.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Shalaby, F. *et al.* A requirement for Flk1 in primitive and definitive hematopoiesis and vasculogenesis. *Cell* **89**, 981–990 (1997).
- Shalaby, F. *et al.* Failure of blood-island formation and vasculogenesis in Flk-1-deficient mice. *Nature* **376**, 62–66 (1995).
- Guo, G. *et al.* Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell* **18**, 675–685 (2010).
- Yan, L. *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
- Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593–597 (2013).
- Pina, C. *et al.* Inferring rules of lineage commitment in haematopoiesis. *Nat. Cell Biol.* **14**, 287–294 (2012).
- Moignard, V. *et al.* Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell Biol.* **15**, 363–372 (2013).
- Swiers, G. *et al.* Early dynamic fate changes in haemogenic endothelium characterized at the single-cell level. *Nat. Commun.* **4**, 2924 (2013).
- Coifman, R.R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci. USA* **102**, 7426–7431 (2005).
- Lux, C.T. *et al.* All primitive and definitive hematopoietic progenitor cells emerging before E10 in the mouse embryo are products of the yolk sac. *Blood* **111**, 3435–3438 (2008).
- Ding, G., Tanaka, Y., Hayashi, M., Nishikawa, S.-I. & Kataoka, H. PDGF receptor  $\alpha$ + mesoderm contributes to endothelial and hematopoietic cells in mice. *Dev. Dyn.* **242**, 254–268 (2013).
- Tanaka, Y. *et al.* Early ontogenic origin of the hematopoietic stem cell lineage. *Proc. Natl. Acad. Sci. USA* **109**, 4515–4520 (2012).
- Lorsbach, R.B. *et al.* Role of RUNX1 in adult hematopoiesis: analysis of RUNX1-IRES-GFP knock-in mice reveals differential lineage expression. *Blood* **103**, 2522–2529 (2004).
- Moignard, V., Woodhouse, S., Fisher, J. & Göttgens, B. Transcriptional hierarchies regulating early blood cell development. *Blood Cells Mol. Dis.* **51**, 239–247 (2013).
- Thiery, J.P., Acloque, H., Huang, R.Y.J. & Nieto, M.A. Epithelial-mesenchymal transitions in development and disease. *Cell* **139**, 871–890 (2009).
- Costa, G., Kouskoff, V. & Lacada, G. Origin of blood cells and HSC production in the embryo. *Trends Immunol.* **33**, 215–223 (2012).
- Samokhvalov, I.M., Samokhvalova, N.I. & Nishikawa, S. Cell tracing shows the contribution of the yolk sac to adult haematopoiesis. *Nature* **446**, 1056–1061 (2007).
- Fujiwara, Y., Browne, C.P., Cuniff, K., Goff, S.C. & Orkin, S.H. Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *Proc. Natl. Acad. Sci. USA* **93**, 12355–12358 (1996).
- Robb, L. *et al.* Absence of yolk sac hematopoiesis from mice with a targeted disruption of the scl gene. *Proc. Natl. Acad. Sci. USA* **92**, 7075–7079 (1995).
- Shivdasani, R.A., Mayer, E.L. & Orkin, S.H. Absence of blood formation in mice lacking the T-cell leukaemia oncogene tal-1/SCL. *Nature* **373**, 432–434 (1995).
- Schlaeger, T.M., Mikkola, H.K.A., Gekas, C., Helgadottir, H.B. & Orkin, S.H. Tie2-Cre-mediated gene ablation defines the stem-cell leukemia gene (SCL/tal1)-dependent window during hematopoietic stem-cell development. *Blood* **105**, 3871–3874 (2005).
- Chen, M.J., Yokomizo, B.M., Zeigler, E., Dzierzak, E. & Speck, N.A. Runx1 is required for the endothelial to haematopoietic cell transition but not thereafter. *Nature* **457**, 887–891 (2009).
- North, T. *et al.* Cbfa2 is required for the formation of intra-aortic hematopoietic clusters. *Development* **126**, 2563–2575 (1999).
- Wareing, S. *et al.* The Flk1-Cre-mediated deletion of ETV2 defines its narrow temporal requirement during embryonic hematopoietic development. *Stem Cells* **30**, 1521–1531 (2012).
- Sumanas, S. *et al.* Interplay among Etsrp/ER71, Scl, and Alk8 signaling controls endothelial and myeloid cell formation. *Blood* **111**, 4500–4510 (2008).
- Krumsiek, J., Marr, C., Schroeder, T. & Theis, F.J. Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PLoS ONE* **6**, e22649 (2011).
- Bonzanni, N. *et al.* Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. *Bioinformatics* **29**, i80–i88 (2013).
- Xu, H., Ang, Y.-S., Sevilla, A., Lemischka, I.R. & Ma'ayan, A. Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. *PLoS Comput. Biol.* **10**, e1003777 (2014).
- Sánchez-Castillo, M. *et al.* CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res.* **43**, D117–D123 (2015).
- Fan, R. *et al.* Dynamic HoxB4-regulatory network during embryonic stem cell differentiation to hematopoietic cells. *Blood* **119**, e139–e147 (2012).
- Thoms, J.A. *et al.* ERG promotes T-acute lymphoblastic leukemia and is transcriptionally regulated in leukemic cells by a stem cell enhancer. *Blood* **117**, 7079–7089 (2011).
- Wilson, N.K. *et al.* The transcriptional program controlled by the stem cell leukemia gene Scl/Tal1 during early embryonic hematopoietic development. *Blood* **113**, 5456–5465 (2009).
- Wilkinson, A.C. *et al.* Single site-specific integration targeting coupled with embryonic stem cell differentiation provides a high-throughput alternative to *in vivo* enhancer analyses. *Biol. Open* **2**, 1229–1238 (2013).
- Mitjavila-Garcia, M.T. *et al.* Expression of CD41 on hematopoietic progenitors derived from embryonic hematopoietic cells. *Development* **129**, 2003–2013 (2002).
- Mikkola, H.K.A., Fujiwara, Y., Schlaeger, T.M., Traver, D. & Orkin, S.H. Expression of CD41 marks the initiation of definitive hematopoiesis in the mouse embryo. *Blood* **101**, 508–516 (2003).
- Kabrun, N. *et al.* Flk-1 expression defines a population of early embryonic hematopoietic precursors. *Development* **124**, 2039–2048 (1997).
- Gandillet, A. *et al.* Sox7-sustained expression alters the balance between proliferation and differentiation of hematopoietic progenitors at the onset of blood specification. *Blood* **114**, 4813–4822 (2009).
- Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A. & Nolan, G.P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–529 (2005).
- Guo, G. *et al.* Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. *Cell Stem Cell* **13**, 492–505 (2013).
- Bailey, N.T.J. *Statistical Methods in Biology* **1995**, 255 (Cambridge University Press, 1995).
- Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
- Pimanda, J.E. *et al.* Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development. *Proc. Natl. Acad. Sci. USA* **104**, 17692–17697 (2007).
- Kataoka, H. *et al.* ETV2/ER71 induces vascular mesoderm from Flk1+PDGFR $\alpha$ + primitive mesoderm. *Blood* **118**, 6975–6986 (2011).
- Loughran, S.J. *et al.* The transcription factor Erg is essential for definitive hematopoiesis and the function of adult hematopoietic stem cells. *Nat. Immunol.* **9**, 810–819 (2008).
- Taoudi, S. *et al.* ERG dependence distinguishes developmental control of hematopoietic stem cell maintenance from hematopoietic specification. *Genes Dev.* **25**, 251–262 (2011).
- Sacilotto, N. *et al.* Analysis of Dll4 regulation reveals a combinatorial role for Sox and Notch in arterial development. *Proc. Natl. Acad. Sci. USA* **110**, 11893–11898 (2013).
- Kyba, M., Perlingeiro, R.C.R. & Daley, G.Q. HoxB4 confers definitive lymphoid-myeloid engraftment potential on embryonic stem cell and yolk sac hematopoietic progenitors. *Cell* **109**, 29–37 (2002).
- Tischler, J. & Surani, M.A. Investigating transcriptional states at single-cell-resolution. *Curr. Opin. Biotechnol.* **24**, 69–78 (2013).
- Tang, F., Lao, K. & Surani, M.A. Development and applications of single-cell transcriptome analysis. *Nat. Methods* **8**, S6–S11 (2011).

## ONLINE METHODS

**Timed matings and embryo collection.** Timed mating were set up between homozygous *Runx1* reporter male and female mice using the *Runx1*-ires-GFP knock-in mouse previously described<sup>13</sup>. Animals also contained a Gata1-mCherry reporter transgene not used in this study. All animal experiments were carried out in accordance with the RIKEN guidelines for animal and recombinant DNA experiments. Embryos were staged according to morphologic criteria<sup>50</sup>. Suspensions of embryo cells were prepared as described previously<sup>12</sup> and single-cell suspensions were stained with Flk-1-APC (AVAS12 at 1:100 dilution; BD Bioscience). Cells were sorted using a FACS Aria II (BD Bioscience) and 100  $\mu$ m nozzle. 4SG cells were not sorted for Flk1 as its expression begins to be downregulated by this time. 4SG<sup>+</sup> cells were specifically *Runx1*-GFP<sup>+</sup> at the protein level in order to exclude committed blood cells of the 4SG population, but could express *Runx1* at the mRNA level.

**Single-cell qRT-PCR.** Single-cell qRT-PCR was carried out using the Fluidigm BioMark platform as described<sup>7</sup>, with a limit of detection (LOD) of Ct 25. The LOD was determined according to Stahlberg *et al.*<sup>51</sup> and manufacturer's instructions. Briefly, standard curves were run on the BioMark with six repeats of each dilution. For each gene, the LOD was the average Ct value for the last dilution at which all six replicates had positive amplification. The overall LOD for the gene set was the median Ct value across all genes. TaqMan assays (Life Technologies) used are listed in **Supplementary Table 1**. Raw Ct values and normalized data can be found in **Supplementary Table 7**. Gene expression was subtracted from the limit of detection and normalized on a cell-wise basis to the mean expression of the four housekeeping genes (*Eif2b1*, *Mrpl19*, *Polr2a* and *Ubc*) in each cell. Cells that did not express all four housekeeping genes were excluded from subsequent analysis, as were cells for which the mean of the four housekeepers was  $\pm 3$  s.d. from the mean of all cells. A dCt value of  $-14$  was then assigned where a gene was not detected. 85–90% of sorted cells were retained for further analysis. *Gata2* did not amplify correctly and *HoxB3* was not expressed in any cells, so these factors have been excluded from the analysis. Further analyses were done on the dCt values for all transcription factors and marker genes, but not housekeeping genes. Hierarchical clustering was performed in R (<http://www.r-project.org/>) using the *hclust* package and displayed with *heatmap.2* from the *gplots* package. Complete linkage was used as recommended by Fluidigm for single-cell expression data, with the distance defined as  $1 - \text{Spearman rank correlation between samples or genes}$ . Optimal ordering of leaves was determined using the *cba* package. Principal components analysis was done using the *prcomp* function in R and PC scores for each cell are shown. Plots were colored retrospectively to indicate the anatomical stage of origin of each cell.

**RNA sequencing.** Cells were sorted into 2  $\mu$ l of lysis buffer (0.2% (v/v) Triton X-100 and 2 U/ $\mu$ l RNase inhibitor (Clontech)) and stored at  $-80^\circ\text{C}$ . RNA-seq was carried out using the Smart-seq2 protocol according to Picelli *et al.*<sup>52</sup> and sequenced on an Illumina HiSeq 2500.

The reads for five biological replicates for each subtype were mapped using the RNA-seq aligner STAR version 2.3.0 (ref. 53). Parameters used to align with STAR were “–outFilterMultimapScoreRange 1–outSAMstrandField intronMotif–genomeLoad NoSharedMemory–outStd SAM”. *Mus musculus* Ensembl assembly GRCm38 (equivalent to UCSC mm10) was used to build the STAR index file, along with the GTF file (version GRCm38.74) from Ensembl. Samtools version 0.1.18 was used to sort the STAR SAM output file and convert it to BAM format, which would then be used as input for the HTSeq-counts program. Counts were determined using the HTSeq-counts program version 0.6.1 (<http://www.huber.embl.de/users/anders/HTSeq/doc/overview.html>) with the parameter “–stranded=no”. FPKM values were calculated using in-house scripts and are provided in **Supplementary Table 8**.

**Diffusion plots.** Our visualization approach is based on the diffusion map formalism<sup>9</sup>. In brief, affinities between all cells based on their expression levels are calculated using a diffusion metric. Next, the cells are organized in 2D or 3D such that the Euclidean distance between the cells corresponds

to the diffusion metric. We determined the cell-cell affinities using an isotropic Gaussian kernel

$$P(i, j) = \frac{1}{Z_i} \exp \left( \frac{-(x_i - x_j)^2}{\epsilon} \right)$$

with  $x_i$  and  $x_j$  being the gene expression vector for cells  $i$  and  $j$  and  $Z_i$  being a normalization constant such that

$$\sum_{i=1}^N P(i, j) = 1$$

The quantity  $P(i, j)$  can then be interpreted as the transition probability of a diffusion process between cells. Consequently, it is particularly well-suited for representing the gradual change in the transcriptional landscape related to developmental trajectories. In contrast, other methods for dimensionality reduction and visualization of high-dimensional data such as t-SNE<sup>54,55</sup>, encourage a representation of the data as disjoint clusters, which is less meaningful for modeling continuous developmental trajectories.

In order to account for the nonuniform density  $\rho$  of cells in the gene expression space (i.e., the potential presence of rare populations), we renormalize the affinities  $P(i, j)$  between two cells  $i$  and  $j$  based on the local density  $\rho(i)$  and  $\rho(j)$  to  $\tilde{P}(i, j)$ . Furthermore, we encourage a better representation of local behavior by using only a subset of neighboring cells (20% nearest neighbors) for computing the affinities and by setting the diagonal of the affinity matrix  $\tilde{P}$  to zero. We then calculate the eigenvectors of  $\tilde{P}$  and retain the eigenvectors with the greatest eigenvalues, which we use for visualization.

**Network analysis.** In computer science, synthesis is a general term for the counterpart of verification. In verification, a hand-built model is given, along with a specification of how it ought to behave. Then the model is checked to ensure it satisfies the specification. In synthesis, a specification is given and a model is automatically generated that satisfies this specification. In biology, the specification is the experimental data that the model should reproduce. In our case, it is the state transition graph, which was derived from the single-cell gene expression data. Synthesis has recently been applied in the context of biology<sup>56</sup>. In that work, state machine-like models were synthesized, which were consistent with known experimental mutation results, given in a genotype-phenotype table. Both the data and the type of model considered were different from those dealt with in the current work, which called for a different approach.

To synthesize a Boolean network model, we would like to orient the transitions in the state graph (previously every pair of states that differ in the expression of exactly one gene were connected by an undirected edge) such that a given set of final states will be reachable from a given set of initial states. We will allow edges to be directed in one direction, both directions, or in neither. We would then like to extract the Boolean update functions that give rise to these directed transitions. We try to get the best possible network by maximizing the number of states in which no transitions induced by the update functions are missing (condition 2, below). We can state our synthesis problem formally as follows. We are given a set of variables  $V$ , corresponding to genes, and an undirected graph  $G = (N, E)$  where each node  $n \in N$  is labeled with a state  $s: V \rightarrow \{0, 1\}$ , and each edge  $\{s_1, s_2\} \in E$  is labeled with the single variable that changes between state  $s_1$  and  $s_2$ . Note that by  $s$  we denote both a state and the node labeled by that state, unambiguously. We are also given a designated set  $I \subseteq N$  of initial vertices and a designated set  $F \subseteq N$  of final vertices, along with a threshold  $t_i$  for each variable  $v_i \in V$ .

Our synthesis method searches for an orientation of  $G$ , along with an update function  $u_i: \{0, 1\}^n \rightarrow \{0, 1\}$  for each variable  $v_i \in V$ , such that the following conditions hold:

1. For each edge  $(s_1, s_2)$  labeled with variable  $v_i$  in the orientated graph, the update function for  $v_i$  takes state  $s_1$  to state  $s_2$ :  $u_i(s_1) = s_2(i)$ .

- For every variable  $v_i \in V$ , let  $N_i$  be the set of states without a  $v_i$ -labeled edge. For every  $i$  the number of states  $s \in N_i$  such that  $u_i(s) = s(i)$  is greater than or equal to  $t_i$ .
- Every final vertex  $f \in F$  is reachable from some initial vertex  $i \in I$  by a directed path in the orientated graph.

We restrict the update function  $u_i$  to have the form:

$$f_1 \wedge \neg f_2$$

where  $f_j$  is a Boolean formula that has and-nodes of in-degree two and/or-nodes of arbitrary in-degree, and where  $f_1$  has a maximum depth of  $N_i$  and  $f_2$  has a maximum depth of  $M_i$ .  $N_i$  and  $M_i$  are given as parameters to the method.

The search for edge orientations and associated Boolean update rules is encoded as a Boolean satisfiability (SAT) problem. The update functions of each variable can be sought after separately, giving rise to reasonably sized satisfiability queries. We then combine compatible single-gene update functions by restricting our attention to combinations that permit paths from initial to final nodes. Paths between initial and final nodes in oriented graphs are found using a breadth-first search for the shortest path between two nodes. We restrict our search to these shortest paths both for efficiency reasons and to eliminate routes that seem “unbiological,” for example, routes that cross a fate transition and then return to where they began. We exhaustively search for all satisfying solutions. The method is implemented in the F# programming language, and uses the Z3 solver to handle SAT queries.

After assessing the method’s capability to reconstruct literature-derived asynchronous Boolean networks from their own state spaces (Supplementary Note), we applied it to our biological data. From the resulting synthesized Boolean network models, we obtained a core network of 20 factors.

For our initial states, we took the set of the PS states in the earliest state cluster in the state transition graph. As the final states, we took a core of the 4SG states in the latest cluster. These states are listed Supplementary Table 9.

Note that owing to the intermixing between populations, there is no guarantee that a state measured on E7.75 is further ahead in development than a state measured on E7.5 (for example). We therefore only constrain reachability from start states to end states, and do not require that experimental measurement time is respected. To obtain the thresholds,  $t_i$ , and the maximum sizes of the activating and repressing portions of update functions  $N$  and  $M$ , we performed an optimization step for each gene independently, where the size of allowed update functions was steadily increased until  $N_i$  reached a maximum.  $t_i$  was then set to  $0.66 N_i$  in order to allow the method room to find Boolean update rules that permit a path from initial to final states. To obtain the stable states of Boolean network models, the algorithm from Garg *et al.*<sup>57</sup> was applied. Binary states can be found in Supplementary Table 10 and cells with equal cell states are listed in Supplementary Table 11.

**Synthesis bootstrapping.** To assess the robustness of the predictions of network synthesis, we performed bootstrapping. A random sample of 75% of the 3,934 gene expression profiles was retained, and a new state transition graph was built from this reduced data set. This state transition graph was then used as the basis to synthesize new Boolean rules, using the same parameters as the original analysis. The results of repeating this process five times are shown in Supplementary Table 3a–e. Bold entries indicate a rule is identical to a rule synthesized from the original data set. Underlined entries indicate that a rule is contained within a larger rule from the original synthesis. We see that in most cases the original rule or a closely related, underlined rule is synthesized. In general, the number of possible solutions for a gene’s update function grows as the amount of data used is decreased, and including the full data set narrows these possibilities.

**Assessing sensitivity of synthesized rules to binary discretization threshold.** In order to construct a state transition graph and apply our synthesis method, experimental data must first be discretized to binary values that indicate whether a gene is expressed or not expressed. The details of how we determine this threshold are covered in the section entitled “Single-cell q-RT-PCR”.

To assess sensitivity of results to the choice of threshold, we repeated our analysis with a more stringent cut off, increasing it by two cycles. This resulted in a state transition graph of 1,249 nodes (199 fewer nodes than the original state transition graph), which was then used as the basis to synthesize new Boolean rules, using the same parameters as the original analysis. The results are shown in Supplementary Table 4. Bold entries indicate a rule is identical to a rule synthesized from the original data set. Underlined entries indicate that a rule is contained within a larger rule from the original synthesis. We see that in most cases the original rule or a closely related, underlined rule is synthesized. In general, the number of possible solutions for a gene’s update function grows as the amount of data used is decreased, and including the full data set narrows these possibilities.

**Erg+85kb enhancer reporter cassette generation.** *Hprt* locus-targeting, enhancer-reporter cassettes containing the *Erg+85kb* (wild type or mutated) element upstream of a Venus YFP fluorescent reporter gene driven by the *Hsp68* minimal promoter (*Erg+85kb/Hsp/Venus*) were generated by Gateway cloning as previously described<sup>33</sup>, and verified by sequencing. The coordinates of the cloned region in the mouse mm10 genome build are chr16:95439106-95439643. The wild-type and mutated *Erg+85kb* elements were initially PCR amplified from synthetic Gene Art Strings (Life Technologies) using primers with attB sequences (underlined) upstream of enhancer-specific sequence (*Erg+85attb1F*

GGGGACAAGTTTGTACAAAAAAGCAGGCTGCCTAAGGGCCGAG GTTG, *Erg+85attb1R* GGGGACCACCTTTGTACAAGAAAGCTGGGTGC ATGAAATC

ACCTTGGAATTTGTC; see Fig. 4a for sequences of mutated motifs).

#### Embryonic stem (ES) cell gene targeting, differentiation and analysis.

*Erg+85kb/Hsp/Venus* cassettes were targeted to the *Hprt* locus in *Hprt*-deficient mouse HM-1 ES cells<sup>58</sup> to generate clonal lines that were differentiated into blood by embryoid body formation and analyzed at the stated time points by flow cytometry for Flk1 (as above) and CD41-PECy7 (eBioMWReg30, 1:500, BioLegend), all as previously described<sup>33</sup>. Data are the combined average of three biological replicates from two ES cell clones. Two HM-1 lines carrying an enhancer-less *Hsp/Venus* cassette were used as a control, as described previously<sup>33</sup>.

**Sox7 induction and colony assays.** Timed matings were set up between transgenic male iSox7<sup>+</sup>rtTA<sup>+</sup> and wild-type ICR female mice<sup>37</sup>. The morning of vaginal plug detection was considered E0.5. All animal work was done under regulations governed by the Home Office Legislation under the Animal Scientific Procedures Act (ASPA) 1986. Cells from E8.25 embryos were tested in a clonogenic replating assay for hematopoietic progenitors with or without 1 µg/ml doxycycline as previously described<sup>37</sup>. For each embryo 1/10 of the cells was used for genotyping and the remaining cells were equally divided into –dox and +dox conditions. Primitive erythroid colonies were counted after 4 days in culture. Primers used for genotyping were rtTA-F ACAAGGTTTCTTCTACTAGAGAACGCG, rtTA-R AGATCGAAATCGT CTAGCGCGTCG, iSox7-F CTAGATCTCGAAGGATCTGGAG, iSox7-R ATACTTTCTCGGCAGGAGCA.

**Availability of computational resources and data.** We provide our SCNS toolkit and associated data at <http://scns.stemcells.cam.ac.uk> as well as in Supplementary Code. This includes the full code for the synthesis method, along with scripts for:

- Constructing a state transition graph from single-cell gene expression data.
- Automating the process of finding stable states and performing all single-gene *in silico* perturbations of synthesized Boolean networks. This second script also categorizes perturbations in terms of alterations to the stable states that the model is able to reach. Both a failure to reach states normally reachable for the wild-type model, as well as stabilization at novel “unnatural” states can be important, with the former mimicking, for example, the failure of a cell to develop down a given lineage, whereas the latter could be used to gain mechanistic understanding of pathological cellular states (such as in cancer cells). For example, to look for factors

involved in blood differentiation, we collected all perturbations which retained the desired “endothelial-like” state, removed the undesired “blood-like” state, and then ranked these perturbations by the number of additional, undesired states that were introduced.

50. Downs, K.M. & Davies, T. Staging of gastrulating mouse embryos by morphological landmarks in the dissecting microscope. *Development* **118**, 1255–1266 (1993).
51. Ståhlberg, A. *et al.* Defining cell populations with single-cell gene expression profiling: correlations and identification of astrocyte subpopulations. *Nucleic Acids Res.* **39**, e24 (2011).
52. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
53. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
54. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
55. Amir, E.-A.D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–552 (2013).
56. Koksai, A.S. *et al.* Synthesis of biological models from mutation experiments. *POPL '13 Proc. 40th Annu. ACM SIGPLAN-SIGACT Symp. Princ. Program. Lang* **48**, 469–482 (2013).
57. Garg, A., Di Cara, A., Xenarios, I., Mendoza, L. & De Micheli, G. Synchronous versus asynchronous modeling of gene regulatory networks. *Bioinformatics* **24**, 1917–1925 (2008).
58. Magin, T.M., McWhir, J. & Melton, D.W. A new mouse embryonic stem cell line with good germ line contribution and gene targeting frequency. *Nucleic Acids Res.* **20**, 3795–3796 (1992).