LETTER

doi:10.1038/nature18633

Resolving early mesoderm diversification through single-cell expression profiling

Antonio Scialdone[1,2]*, Yosuke Tanaka[3,4]*†, Wajid Jawaid[3,4]*, Victoria Moignard[3,4]*, Nicola K. Wilson[3,4], Iain C. Macaulay[2], John C. Marioni[1,2,5] & Berthold Göttgens[3,4]

In mammals, specification of the three major germ layers occurs during gastrulation, when cells ingressing through the primitive streak differentiate into the precursor cells of major organ systems. However, the molecular mechanisms underlying this process remain unclear, as numbers of gastrulating cells are very limited. In the mouse embryo at embryonic day 6.5, cells located at the junction between the extra-embryonic region and the epiblast on the posterior side of the embryo undergo an epithelial-to-mesenchymal transition and ingress through the primitive streak. Subsequently, cells migrate, either surrounding the prospective ectoderm contributing to the embryo proper, or into the extra-embryonic region to form the yolk sac, umbilical cord and placenta. Fate mapping has shown that mature tissues such as blood and heart originate from specific regions of the pre-gastrula epiblast[1], but the plasticity of cells within the embryo and the function of key cell-type-specific transcription factors remain unclear. Here we analyse 1,205 cells from the epiblast and nascent Flk1$^+$ mesoderm of gastrulating mouse embryos using single-cell RNA sequencing, representing the first transcriptome-wide *in vivo* view of early mesoderm formation during mammalian gastrulation. Additionally, using knockout mice, we study the function of Tal1, a key haematopoietic transcription factor, and demonstrate, contrary to previous studies performed using retrospective assays[2,3], that *Tal1* knockout does not immediately bias precursor cells towards a cardiac fate.

Traditional experimental approaches for genome-scale analysis rely on large numbers of input cells and therefore cannot be applied to study early lineage diversification directly in the embryo. To address this, we used single-cell transcriptomics to investigate mesodermal lineage diversification towards the haematopoietic system in 1,205 single cells covering a time course from early gastrulation at embryonic day (E)6.5 to the generation of primitive red blood cells at E7.75 (Fig. 1a and Extended Data Figs 1a and 2a). Using previously published metrics (Methods), we observed that the data were of high quality. Five hundred and one single-cell transcriptomes were obtained from cells taken from dissected distal halves of E6.5 embryos sorted for viability only, which contain all of the epiblast cells, including the developing primitive streak, and a limited number of visceral endoderm and extra-embryonic ectoderm cells. From E7.0, embryos were staged according to anatomical features (Methods) as primitive streak, neural plate and head fold. The VEGF receptor Flk1 (*Kdr*) was used to capture cells as it marks much of the developing mesoderm[4]. During subsequent blood development, Flk1 is downregulated and CD41 (*Itga2b*) is upregulated[5]. We therefore also sampled cells expressing both markers and CD41 alone at the neural plate and head fold stages (Fig. 1a and Extended Data Figs 1b and 2a), giving a total of 138 cells from E7.0 (primitive streak), 259 from E7.5 (neural plate) and 307 from E7.75 (head fold).

After rigorous quality control, 2,085 genes were identified as having significantly more heterogeneous expression across the 1,205 cells than expected by chance (Extended Data Fig. 2b–d). Unsupervised hierarchical clustering in conjunction with a dynamic hybrid cut (Methods)

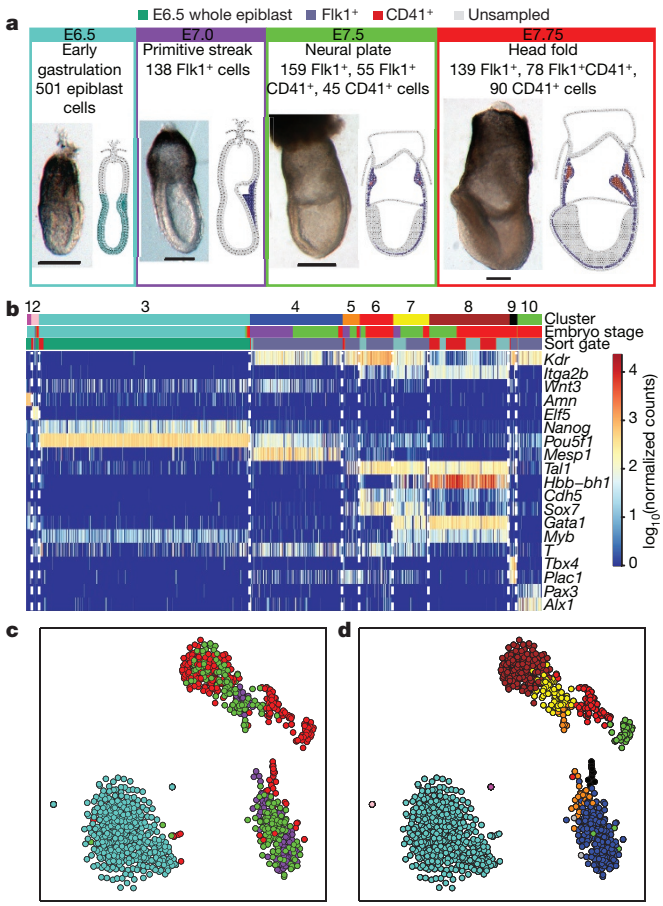


Figure 1 | Single-cell transcriptomics identifies ten populations relevant to early mesodermal development. a, Whole-mount images and schematics of E6.5–7.75 embryo sections. Colours indicate approximate locations of sorted cells. Anterior, left; posterior, right. Scale bars, 200 μm. **b**, Heatmap showing key genes distinguishing ten clusters. Coloured bars indicate assigned cluster (top), stage (middle: turquoise, E6.5; purple, primitive streak (E7.0); green, neural plate (E7.5); red, head fold (E7.75)) and the sorted population (bottom: green, E6.5 epiblast; blue, Flk1$^+$; turquoise, Flk1$^+$CD41$^+$; red, Flk1$^-$CD41$^+$). **c**, t-SNE of all 1,205 cells coloured by embryonic stage, and (**d**) according to clusters in **b**.

[1]EMBL-European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge CB10 1SD, UK. [2]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge CB10 1SA, UK. [3]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK. [4]Wellcome Trust - Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. [5]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, UK. †Present address: Division of Cellular Therapy, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.
*These authors contributed equally to this work.

yielded ten robust clusters with varying contributions from the different embryonic stages (Fig. 1b, Extended Data Fig. 3, Methods and cell numbers in Extended Data Fig. 3h). Using t-distributed stochastic neighbour embedding (t-SNE) dimensionality reduction to visualize the data, three major groups were observed: one comprising almost all E6.5 cells, another mainly consisting of earlier primitive streak and neural plate stage cells, and a third containing predominantly later head fold stage cells (Fig. 1c). Importantly, clusters were coherent with the t-SNE visualization except for the small cluster 5 (Fig. 1d).

The expression of key marker genes allowed us to assign identities to each cluster: visceral endoderm, extra-embryonic ectoderm, epiblast, early mesodermal progenitors, posterior mesoderm, endothelium, blood progenitors, primitive erythrocytes, allantoic mesoderm and pharyngeal mesoderm (Fig. 1b, Extended Data Figs 3h and 4). Because of the limited cell numbers and lack of markers for their prospective isolation, conventional bulk transcriptome analysis of these key populations has never before been attempted.

Since the T-box transcription factor Brachyury—encoded by the *T* gene—marks the nascent primitive streak[6], we investigated the gene expression programs associated with *T* induction in the E6.5 cells (cluster 3). *T* expression was restricted to a distinct subset of epiblast cells found closest to cluster 4 (Fig. 1d and Extended Data Fig. 5b), with rare isolated cells within the bulk of the epiblast population also expressing moderate levels, consistent with priming events for single gastrulation-associated genes. *T* expression correlated with other gastrulation-associated genes including *Mixl1* and *Mesp1* (Fig. 2a), with *Mesp1* highly expressed only in the small subset of cells situated at the pole of the E6.5 epiblast cluster (association of *T* and *Mesp1* expression: *P* value 3×10^{-15}, Fisher's exact test). We also observed a subset of cells distinct from the $T^{+}/Mesp1^{+}$ population, which expressed *Foxa2*, suggestive of endodermal priming[7] (Extended Data Fig. 5d).

We next identified genes displaying correlated expression with *T*, which identified known markers and regulators such as *Mixl1*, and genes not previously implicated in mammalian gastrulation, such as *Slc35d3*, an orphan member of a nucleotide sugar transporter

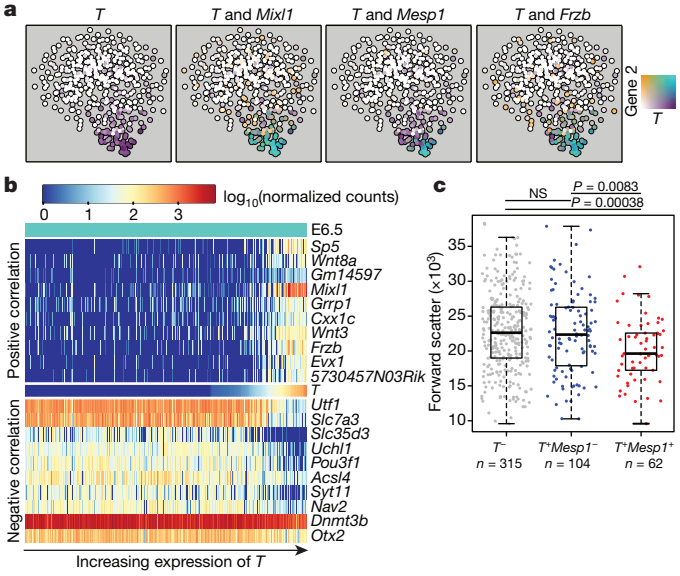


Figure 2 | Transcriptional program associated with *T* induction in E6.5 epiblast cells. a, t-SNE of the 481 E6.5 cells in cluster 3. Points are coloured by expression of *T* (Brachyury) and *Mixl1*, *Mesp1* and *Frzb*. **b**, Heatmap showing the ten genes most highly positively and negatively correlated with *T* (Supplementary Information Table 1). **c**, Forward scatter for the 481 E6.5 epiblast cells in cluster 3, with cells grouped according to *T*/*Mesp1* expression. Boxplots indicate the median and interquartile range. *P* values were calculated using a two-sided Welch's *t*-test for samples with unequal variance, with false discovery rate correction for multiple testing.

family[8] and the retrotransposon-derived transcript *Cxx1c*[9] (Fig. 2b and Supplementary Information Table 1). Genes negatively correlated with *T* were consistently expressed across the majority of epiblast cells, suggesting that cells outside the primitive streak have not yet committed to a particular fate, consistent with the known plasticity of epiblast cells in transplant experiments[10]. Ingressing epiblast cells undergo an EMT, turning from pseudo-stratified epithelial cells into individual motile cells, a conformational change associated with alterations in cell size and shape[11]. Our E6.5 epiblast cells were isolated using index sorting, thus providing a forward scatter value for each cell. As shown in Fig. 2c, $T^{+}/Mesp1^{+}$ co-expressing cells showed a significant reduction in forward scatter values compared with $T^{+}/Mesp1^{-}$ and T^{-} cells. Since forward scatter correlates positively with cell size, this observation provides a direct link between specific transcriptional programs and characteristic physical changes associated with gastrulation. As $T^{+}/Mesp1^{+}$ cells also express *Mesp2*, this observation was consistent with the known EMT defect in *Mesp1*/*Mesp2* double knockout embryos[12]. Index sorting therefore linked expression changes with dynamic physical changes similar to those recognized to occur during chicken gastrulation[13].

We next focused on mesodermal lineage divergence during and immediately after gastrulation. We reasoned that approaches analogous to those used to order single cells in developmental pseudotime could be used to infer the location of cells in pseudospace, specifically with respect to the anterior–posterior axis of the primitive streak (Fig. 3a). To this end, we used diffusion maps[14], a dimensionality reduction technique particularly suitable for reconstructing developmental trajectories[15]. We identified the diffusion-space direction that most probably represents true biological effects (see Methods), which we interpreted as the pseudospace coordinate (red line in Fig. 3b and Extended Data Fig. 6a–d). Hierarchical clustering revealed three groups of genes (Fig. 3c, Extended Data Fig. 6e and Supplementary Information Table 4) showing a gradient of expression along the pseudospace axis. These were assigned as anterior (darker blue, 334 genes) and posterior (lighter blue, 87 genes) owing to the enrichment of genes with known differential expression along the anterior–posterior axis of the primitive streak (Fig. 3d and Extended Data Figs 6f–h and 7). A third cluster was expressed highly at either end of the pseudospace axis (turquoise, 41 genes). Interestingly, the more posterior $Flk1^{+}$ mesodermal cells are associated with the allantois, blood and endothelial clusters (Fig. 1d and Extended Data Fig. 5c), which are known to arise from the posterior primitive streak. Gene ontology analysis revealed that the putative anterior genes were associated with terms relating to somite development, endoderm development and Notch signalling, consistent with a more anterior mesoderm identity[16] (Supplementary Information Table 2a and Extended Data Fig. 6h). Conversely, the putative posterior mesoderm cluster was associated with BMP signalling, hindlimb development and endothelial cell differentiation, consistent with the posterior portion of the streak[17].

Although derived from the same embryonic stages as the mesodermal progenitor cells, cluster 7 lacks expression of genes such as *Mesp1*, yet expresses *Tal1*, *Sox7*, *Tek* (Tie2) and *Fli1*, which are vital for extra-embryonic mesoderm formation (Fig. 1b and Extended Data Fig. 5, 7). Expression of *Kdr* and *Itga2b* (Extended Data Fig. 5b) further highlights clusters 7 and 8 (brown) as corresponding to the developmental journey towards blood, with a transition to mostly head fold stage cells in cluster 8 and increasing expression of embryonic haemoglobin *Hbb-bh1* (Fig. 1b). Given the apparent trajectory of blood development from cluster 7 to 8, we used an analogous approach to that described above to recover a pseudotemporal ordering of cells (Fig. 4a, Extended Data Fig. 8a–d and Methods). Eight hundred and three genes were downregulated, including the haematovascular transcription factor *Sox7*, which is known to be downregulated during blood commitment[15] (Fig. 4c, d and Extended Data Fig. 8e, f). Sixty-seven genes were upregulated including the erythroid-specific transcription factors *Gata1* and *Nfe2*, and embryonic globin *Hbb-bh1*

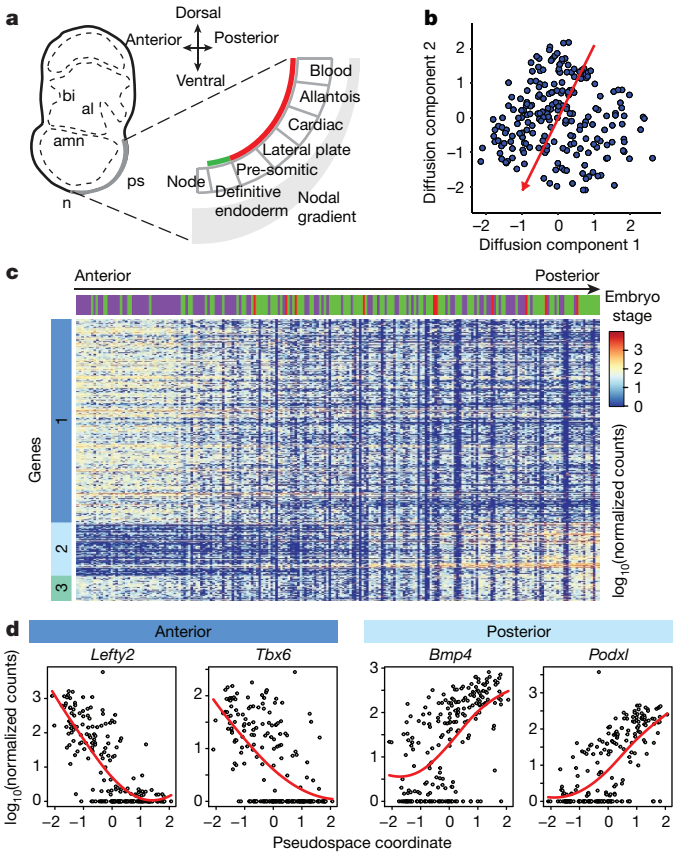


Figure 3 | Dimensionality reduction reveals transcriptional profiles associated with cell location in the embryo. a, Schematic of tissue emergence along the anterior–posterior primitive streak, derived from ref. 29. Mesodermally and endodermally derived tissues are marked by a red and green line, respectively; bi, blood island; al, alantois; amn, amnion; ps, primitive streak; n, node. **b**, Diffusion map of 216 cells in cluster 4 with pseudospace axis in red. Projections onto this axis represent pseudospace coordinates. **c**, Heatmap for differentially expressed genes along the pseudospace axis, showing genes more highly expressed in the anterior (dark blue) and posterior region (light blue), or highly expressed at either end (aquamarine). **d**, Expression profiles for example genes (red line, local polynomial fit).

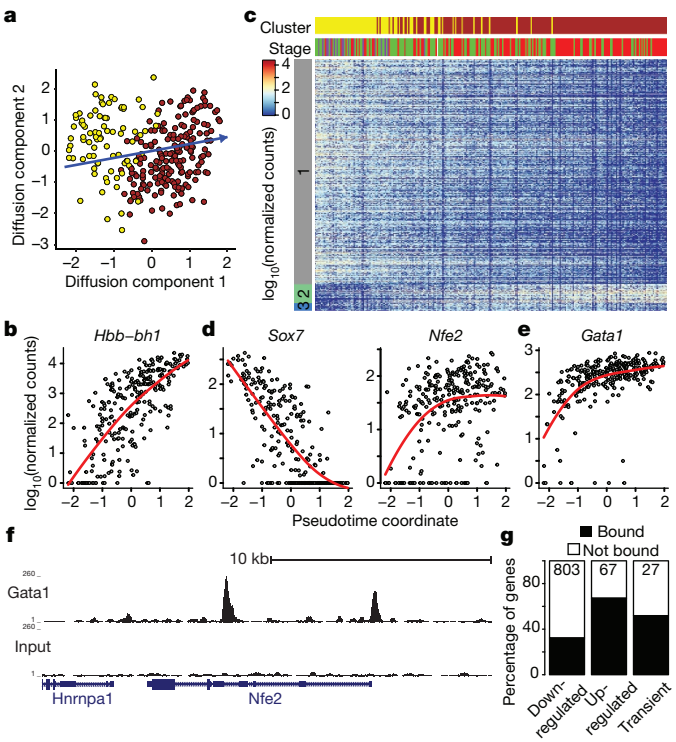


Figure 4 | Inferring the transcriptional program underlying primitive erythropoiesis. a, Diffusion map of 271 cells in clusters 7 and 8 displaying the inferred pseudotime axis (blue). **b**, Expression of *Hbb-bh1* ordered by pseudotime (red line, local polynomial fit). **c**, Heatmap ordered along the pseudotime axis. Horizontal bars indicate cluster and developmental stage. Genes shown were repressed (grey), activated (green) or transiently expressed (blue). **d**, Examples of activated and repressed genes and (**e**) *Gata1* as in **b**. **f**, University of California, Santa Cruz Browser tracks for Gata1 ChIP-seq and input in Runx1^{+}Gata1^{+} cells; the *Nfe2* locus is shown. **g**, Percentage of genes in each group identified in **c** overlapping Gata1 targets. Numbers indicate total numbers of genes in each category from **c**.

(Fig. 4b, d, e and Extended Data Fig. 8). Twenty-seven genes were transiently expressed, including the known erythroid regulator *Gfi1b* (Supplementary Information Table 5). Significant GO terms associated with the upregulated genes were indicative of erythroid development, while downregulated genes were associated with other mesodermal processes including vasculogenesis and osteoblast differentiation (Supplementary Information Table 2b).

Gata1-null embryos die at around E10.5 owing to the arrest of yolk sac erythropoiesis[18]. We generated genome-wide ChIP-seq (chromatin immunoprecipitation followed by sequencing) data for Gata1 in haematopoietic cells derived after 5 days of embryonic stem cell (ESC) *in vitro* differentiation (Extended Data Fig. 9a–c). The group of upregulated genes from the pseudotime analysis showed a pronounced overlap with Gata1 targets ($P < 2.2 \times 10^{-16}$, Fisher's test) including known targets such as *Nfe2* and *Zfpm1* (Fig. 4f, g, Extended Data Fig. 9d, e and Supplementary Information Table 6). Integration of single-cell transcriptomics with complementary transcription factor binding data therefore predicts likely *in vivo* targets of developmental regulators such as Gata1.

Two contrasting mechanisms are commonly invoked to explain how drivers of cell fate determination regulate cell type diversification. The first involves fate restriction through a stepwise sequence of binary fate choices and is supported by mechanistic investigations using ESC differentiation[2,19]. The alternative invokes acquisition of diverse fates from independent precursor cells and is commonly supported by cell transplantation and lineage tracing analysis (Fig. 5a)[1,10,20,21]. In contrast to the retrospective nature of transplantation and lineage tracing experiments where measurements are typically obtained a day or more after cell fate decisions are made, single-cell transcriptomics allows cellular states to be determined at the moment when fate decisions are executed since low cell numbers are not a limiting factor.

The bHLH transcription factor Tal1 (also known as Scl) is essential for the development of all blood cells[22,23] with strong expression in posterior mesodermal derivatives (Fig. 5b). *Tal1*$^{-/-}$ bipotential blood/endothelial progenitors cannot progress to a haemogenic endothelial state[19], *Tal1* overexpression drives transdifferentiation of fibroblasts into blood progenitors[24] and *Tal1*$^{-/-}$ mesodermal progenitors from the yolk sac give rise to aberrant cardiomyocyte progenitors when cultured *in vitro*[2]. However, the precise nature of the molecular defect within *Tal1*$^{-/-}$ mesodermal progenitors within the embryo has remained obscure, because cell numbers are too small for conventional analysis.

We profiled single Flk1^{+} cells from 4 wild type (WT) and 4 *Tal1*$^{-/-}$ embryos obtained from E7.5 (neural plate) to E8.25 (four-somite stage) (256 WT and 121 *Tal1*$^{-/-}$ cells; Fig. 5c and Extended Data Fig. 10), and computationally assigned cells to the previously defined 10 clusters (Methods). Cells from WT embryos contributed to all clusters, while *Tal1*$^{-/-}$ embryos did not contain any cells corresponding to the blood progenitor and primitive erythroid clusters (yellow and brown, Fig. 5d) consistent with the known failure of primitive erythropoiesis in *Tal1*$^{-/-}$ embryos[23] and their lack of CD41 expression (Fig. 5c).

Forty-five *Tal1*$^{-/-}$ cells were confidently mapped to the endothelial (red) cluster, which therefore allowed us to investigate the early consequences of *Tal1* deletion in this key population for definitive

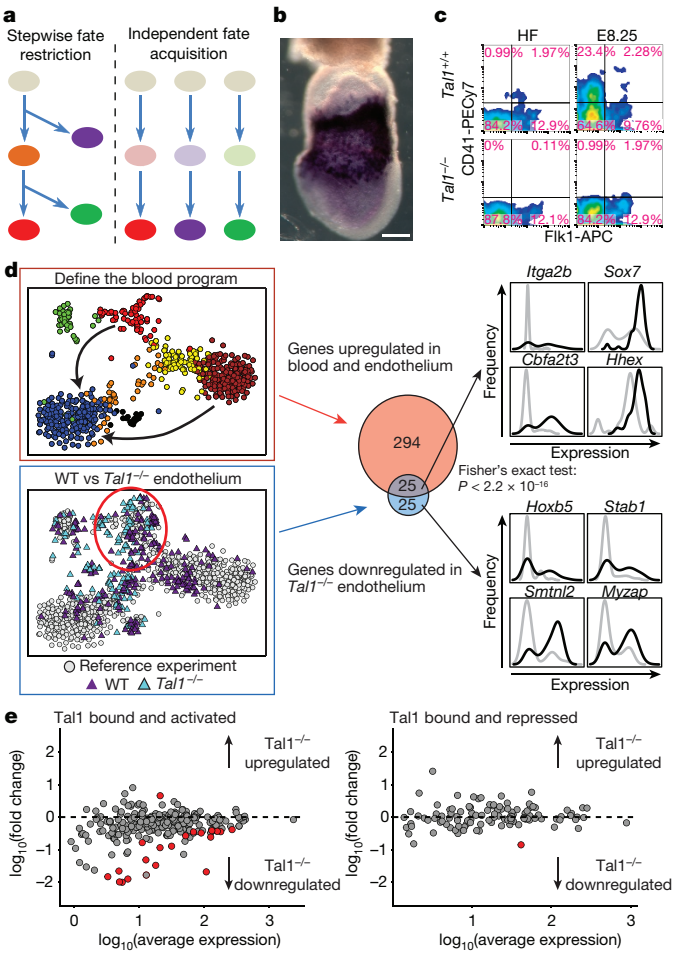


Figure 5 | Analysis of *Tal1*$^{-/-}$ embryos suggests independent fate acquisition. a, Two cell fate diversification models. **b**, *Tal1 in situ* hybridization at head fold stage. Scale bar, 200 μm. **c**, Flow cytometry of WT and *Tal1*$^{-/-}$ mice at head fold and E8.25. **d**, Blood program genes are differentially expressed between nascent mesoderm (blue) and endothelial (red) and blood cells (brown). Differential expression between 45 *Tal1*$^{-/-}$ and 59 WT endothelial cells (lower left t-SNE) identified 50 downregulated genes. Gene set overlap (centre) indicates failure to induce the blood program in *Tal1*$^{-/-}$ endothelium ($P < 2.2 \times 10^{-16}$, Fisher's test). On the right are expression distributions for selected genes in WT (black) or *Tal1*$^{-/-}$ (grey) endothelial cells. **e**, For genes previously reported[3] to be bound and activated (left) or bound and repressed (right) by Tal1, fold change between *Tal1*$^{-/-}$ and WT endothelium (defined in **d**) is plotted against average expression. Red circles, genes with a fold change >1.5 and a false discovery rate <0.05.

haematopoietic development (Fig. 5d and Supplementary Information Tables 7 and 8). Fifty genes were downregulated in *Tal1*$^{-/-}$ endothelial cells (fold change < 0.67, 5% false discovery rate). These included known regulators of early blood development (*Itga2b*, *Lyl1*, *Cbfa2t3*, *Hhex*, *Fli1*, *Ets2*, *Egfl7*, *Sox7*, *Hoxb5*), consistent with Tal1 specifying a haematopoietic fate in embryonic endothelial progenitor cells[19], and in particular *Hoxb5*, which has recently emerged as a powerful marker for definitive blood stem cells[25]. Single-cell profiling also identified genes with altered distributions of expression. For example, *Sox7* changed from a largely unimodal pattern in WT cells to a bimodal on/off pattern in *Tal1*$^{-/-}$ endothelial cells, while *Cbfa2t3* showed the opposite pattern (Fig. 5d).

However, we did not observe upregulation of cardiac markers in *Tal1*$^{-/-}$ endothelial cells (Fig. 5e and Supplementary Information Tables 8 and 9). Previously, this upregulation had been observed in yolk sac endothelial cells collected 1–1.5 days later than our data[2], and had been taken as evidence that Tal1 acts as a gatekeeper controlling the balance between alternative cardiac and blood/endothelial fates within single multipotent mesodermal progenitors[3]. Our results, however, suggest that the primary role of Tal1 is induction of a blood program, and the subsequent ectopic expression of cardiac genes may be the result of secondary induction events acting on a still relatively plastic mesodermal cell blocked from executing its natural developmental program.

Here we have used single-cell transcriptomics to obtain a comprehensive view of the transcriptional programs associated with mammalian gastrulation and early mesodermal lineage diversification. Further technological advances to resolve epigenetic processes at single-cell resolution[26] and match single-cell expression profiles with spatial resolution[27,28] are probably key drivers of future progress in this field. Finally, our analysis of *Tal1*$^{-/-}$ embryos illustrates how the phenotypes of key regulators can be re-evaluated at single-cell resolution to advance our understanding of early mammalian development.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 March; accepted 9 June 2016.
Published online 6 July 2016.

1. Lawson, K. A., Meneses, J. J. & Pedersen, R. A. Clonal analysis of epiblast fate during germ layer formation in the mouse embryo. *Development* **113,** 891–911 (1991).
2. Van Handel, B. *et al.* Scl represses cardiomyogenesis in prospective hemogenic endothelium and endocardium. *Cell* **150,** 590–605 (2012).
3. Org, T. *et al.* Scl binds to primed enhancers in mesoderm to regulate hematopoietic and cardiac fate divergence. *EMBO J.* **34,** 759–777 (2015).
4. Ema, M. *et al.* Primitive erythropoiesis from mesodermal precursors expressing VE-cadherin, PECAM-1, Tie2, endoglin, and CD34 in the mouse embryo. *Blood* **108,** 4018–4024 (2006).
5. 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).
6. Wilkinson, D. G., Bhatt, S. & Herrmann, B. G. Expression pattern of the mouse T gene and its role in mesoderm formation. *Nature* **343,** 657–659 (1990).
7. Burtscher, I. & Lickert, H. Foxa2 regulates polarity and epithelialization in the endoderm germ layer of the mouse embryo. *Development* **136,** 1029–1038 (2009).
8. Chintala, S. *et al.* The Slc35d3 gene, encoding an orphan nucleotide sugar transporter, regulates platelet-dense granules. *Blood* **109,** 1533–1540 (2007).
9. Henke, C. *et al.* Selective expression of sense and antisense transcripts of the sushi-ichi-related retrotransposon – derived family during mouse placentogenesis. *Retrovirology* **12,** 9 (2015).
10. Tam, P. P. L. & Zhou, S. X. The allocation of epiblast cells to ectodermal and germ-line lineages is influenced by the position of the cells in the gastrulating mouse embryo. *Dev. Biol.* **178,** 124–132 (1996).
11. Solnica-Krezel, L. & Sepich, D. S. Gastrulation: making and shaping germ layers. *Annu. Rev. Cell Dev. Biol.* **28,** 687–717 (2012).
12. Kitajima, S., Takagi, A., Inoue, T. & Saga, Y. MesP1 and MesP2 are essential for the development of cardiac mesoderm. *Development* **127,** 3215–3226 (2000).
13. Rozbicki, E. *et al.* Myosin-II-mediated cell shape changes and cell intercalation contribute to primitive streak formation. *Nature Cell Biol.* **17,** 397–408 (2015).
14. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31,** 2989–2998 (2015).
15. Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnol.* **33,** 269–276 (2015).
16. Saga, Y. Segmental border is defined by the key transcription factor Mesp2, by means of the suppression of Notch activity. *Dev. Dyn.* **236,** 1450–1455 (2007).
17. Lawson, K. A. *et al.* Bmp4 is required for the generation of primordial germ cells in the mouse embryo. *Genes Dev.* **13,** 424–436 (1999).
18. Fujiwara, Y., Browne, C. P., Cunniff, 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).
19. Lancrin, C. *et al.* The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage. *Nature* **457,** 892–895 (2009).
20. Padrón-Barthe, L. *et al.* Clonal analysis identifies hemogenic endothelium and not hemangioblasts as the source of the blood-endothelial common lineage in the mouse embryo. *Blood* **124,** 2523–2532 (2014).
21. Tam, P. P., Parameswaran, M., Kinder, S. J. & Weinberger, R. P. The allocation of epiblast cells to the embryonic heart and other mesodermal lineages: the role of ingression and tissue movement during gastrulation. *Development* **124,** 1631–1642 (1997).
22. Porcher, C. *et al.* The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell* **86,** 47–57 (1996).

23. Shivdasani, R. A., Mayer, E. L. & Orkin, S. H. Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature* **373,** 432–434 (1995).
24. Batta, K., Florkowska, M., Kouskoff, V. & Lacaud, G. Direct reprogramming of murine fibroblasts to hematopoietic progenitor cells. *Cell Reports* **9,** 1871–1884 (2014).
25. Chen, J. Y. *et al.* Hoxb5 marks long-term haematopoietic stem cells and reveals a homogenous perivascular niche. *Nature* **530,** 223–227 (2016).
26. Bheda, P. & Schneider, R. Epigenetics reloaded: the single-cell revolution. *Trends Cell Biol.* **24,** 712–723 (2014).
27. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature Biotechnol.* **33,** 503–509 (2015).
28. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnol.* **33,** 495–502 (2015).
29. Robertson, E. J. Dose-dependent Nodal/Smad signals pattern the early mouse embryo. *Semin. Cell Dev. Biol.* **32,** 73–79 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. de Bruijn, A. Martinez-Arias, J. Nichols and C. Mulas for discussion, the Cambridge Institute for Medical Research Flow Cytometry facility for their expertise in single-cell index sorting, and S. Lorenz from the Sanger Single Cell Genomics Core for supervising purification of *Tal1*$^{-/-}$ sequencing libraries. ChIP-seq reads were processed by R. Hannah. Research in the authors' laboratories is supported by the Medical Research Council, Cancer Research UK, the Biotechnology and Biological Sciences Research Council, Bloodwise, the Leukemia and Lymphoma Society, and the Sanger-EBI Single Cell Centre, and by core support grants from the Wellcome Trust to the Cambridge Institute for Medical Research and Wellcome Trust - MRC Cambridge Stem Cell Institute and by core funding from Cancer Research UK and the European Molecular Biology Laboratory. Y.T. was supported by a fellowship from the Japan Society for the Promotion of Science. W.J. is a Wellcome Trust Clinical Research Fellow. A.S. is supported by the Sanger-EBI Single Cell Centre. This work was funded as part of Wellcome Trust Strategic Award 105031/D/14/Z 'Tracing early mammalian lineage decisions by single-cell genomics' awarded to W. Reik, S. Teichmann, J. Nichols, B. Simons, T. Voet, S. Srinivas, L. Vallier, B. Göttgens and J. Marioni.

Author Contributions A.S. and W.J. processed and analysed single-cell RNA sequencing (RNA-seq) data. A.S. and V.M. generated figures. Y.T. and W.J. performed embryo dissection. N.K.W., V.M. and I.C.M. performed single-cell RNA-seq experiments. Y.T. performed flow cytometry, ESC differentiation and *in situ* hybridization. V.M. performed ChIP-seq assays. A.S., W.J., Y.T., V.M., B.G. and J.C.M. interpreted results and wrote the paper. B.G. and J.C.M. supervised and conceived the study.

Author Information ChIP-seq data are available at the NCBI Gene Expression Omnibus portal under accession number GSE74994. Processed data are also available at http://codex.stemcells.cam.ac.uk. RNAseq data are available at Array Express under accession numbers E-MTAB-4079 and E-MTAB-4026. Processed RNAseq data are also available at http://gastrulation.stemcells.cam.ac.uk/scialdone2016. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.G. (bg200@cam.ac.uk) or J.C.M. (marioni@ebi.ac.uk).

Reviewer Information *Nature* thanks A.-K. Hadjantonakis, P. Robson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Timed matings and embryo collection. All procedures were performed in strict adherence to United Kingdom Home Office regulations (project licence 70/8406). Timed matings were set up between CD1 mice (which produce large litters). Embryos were staged according to the morphological criteria of Downs and Davies[30], and classified broadly as primitive streak, neural plate or head fold stage. Suspensions of cells from individual embryos were prepared by incubating with TrypLE Express dissociation reagent (Life Technologies) at 37 °C for 10 min and quenching with heat-inactivated serum. All cells were stained with DAPI for viability. At E6.5, the distal half of the embryo was dissected and dissociated into a single-cell suspension, and live cells were sorted. For E7.0 and older, suspensions consisted of the whole embryo and were also stained with Flk1-APC (AVAS12 at 1:400 dilution; BD Bioscience) and only Flk1$^+$ cells were collected. For cell sorting of CD41$^+$Flk1$^-$ and CD41$^+$Flk1$^+$ cells from neural plate stage and head fold stages, suspensions were stained with Flk1-APC, PDGFRa-PE (APA5 at 1:200 dilution; Biolegend) and CD41-PEcy7 (MWReg30 at 1:400 dilution; Biolegend) for 20 min at 4 °C as described[31]. Cells were sorted from seven E6.5 embryos. Flk1$^+$ cells were sorted from three primitive streak stage, four neural plate stage and three head fold stage embryos (Extended Data Fig. 1a). CD41$^+$Flk1$^-$ and CD41$^+$Flk1$^+$ cells were sorted from the same embryos, an additional eight each at neural plate and head fold stages (Extended Data Fig. 1b). Cell sorting was performed with a BD Influx cell sorter in single-cell sort mode with index sorting to confirm the presence of a single event in each well. Additional cells were sorted into tissue culture plates to visually confirm the presence of single events.

To obtain *Tal1*$^{-/-}$ cells, timed matings were set up between *Tal1*$^{LacZ/+}$ mice[32]. Flk1$^+$ cells were sorted as above from four embryos for each genotype: from one embryo for each genotype at neural plate and four-somite (4S) stages, from two head fold stage embryos for *Tal1*$^{LacZ/LacZ}$ (designated *Tal1*$^{-/-}$), one head fold stage WT embryo and one WT embryo intermediate between neural plate and head fold stages. Genotyping PCR using 1/20 suspension cells was performed as described previously[32].

Single-cell RNA sequencing library preparation and mapping of reads. scRNA-seq analysis used the Smart-seq2 protocol as previously described[33]. Single cells were sorted by fluorescence-activated cell sorting (FACS) into individual wells of a 96-well plate containing lysis buffer (0.2% (v/v) Triton X-100 and 2 U/μl RNase inhibitor (Clontech)) and stored at −80 °C. Libraries were prepared using the Illumina Nextera XT DNA preparation kit and pooled libraries of 96 cells were sequenced on the Illumina Hi-Seq 2500. Reads were mapped simultaneously to the *Mus musculus* genome (Ensembl version 38.77) and the ERCC sequences using GSNAP (version 2014-10-07) with default parameters. HTseq-count[34] was used to count the number of reads mapped to each gene (default options).

Identification of poor quality cells. To assess data quality[35], five metrics were used: (1) total number of mapped reads, (2) fraction of total reads mapped to endogenous genes, (3) fraction of reads mapped to endogenous genes that are allocated to mitochondrial genes, (4) fraction of total reads mapped to ERCC spike-ins and (5) level of sequence duplication (as estimated by FastQC, version 0.11.4, http://www.bioinformatics.babraham.ac.uk/projects/fastqc).

For all downstream analyses, we only retained samples that had (1) more than 200,000 reads mapped (either to ERCC spike-ins or endogenous mRNA), (2) more than 20% of total reads mapped to mRNA, (3) less than 20% of mapped reads allocated to mitochondrial genes, (4) less than 20% of reads mapped to ERCC spike-ins and (5) less than 80% of duplicated sequences. Out of the 2,208 cells that were captured across the two experiments, 1,582 (that is, ∼72% of the total) passed our quality check. A t-SNE projection[36] of the values of these five metrics (Extended Data Fig. 2b) shows that most of discarded cells tend to cluster together and fail at least two criteria. All metrics were standardized before applying t-SNE with the 'RtSNE' function (default parameters) from the R package 'RtSNE' (version 0.1)[37].

Normalization of read counts. The data were normalized for sequencing depth using size factors[38] calculated on endogenous genes. By doing so, we also normalized for the amount of RNA obtained from each cell[39], which is itself highly correlated with cell cycle stage[40].

Highly variable genes and GO enrichment analysis. Highly variable genes were identified by using the method described in Brennecke *et al.*[39]. In brief, we fitted the squared coefficient of variation as function of the mean normalized counts[39]. In the fitting procedure, to minimize the skewing effect due to the lowly expressed genes[39], only genes with a mean normalized count greater than 10 were used. Genes with an adjusted *P* value (Benjamini–Hochberg method) less than 0.1 were considered significant (red circles in Extended Data Fig. 2c). This set of highly variable genes was used for the clustering analysis discussed below. The GO enrichment analysis was performed using TopGO in its 'elimination mode' with Fisher's exact test; we considered as significant GO categories with an unadjusted *P* value below 10^{-4}.

Differentially expressed genes. To find genes differentially expressed between two groups of cells we used edgeR[41] (version 3.12). Before running edgeR, we excluded genes annotated as pseudogenes in Ensembl, sex-related genes (Xist and genes on the Y chromosome) and genes that were not detected or were expressed at very low levels (we considered only genes that had more than ten reads per million in at least *n* cells, *n* being equal to 10% of the cells in the smaller group being compared). The function 'glmTreat' was then used to identify the genes having a fold change significantly greater than 1.5 at a false discovery rate threshold equal to 0.05.

Clustering analysis. Clustering analysis was performed on the 1,205 WT cells from the first experiment that passed the QC. The Spearman correlation coefficient, ρ, was computed between the expression levels of highly variable genes in each pair of cells, which was then used to build a dissimilarity matrix defined as $(1-\rho)/2$. Hierarchical clustering was performed ('hclust' R function with the 'average' method) on the dissimilarity matrix and clusters were identified by means of the dynamic hybrid cut algorithm[42]. The R function 'cutreeDynamic' with the 'hybrid' method and a minimum cluster size equal to ten cells was used ('dynamicTreeCut' package, version 1.62). This function allows the user to specify the 'deepSplit' parameter that controls the sensitivity of the method: higher values of this parameter correspond to higher sensitivity and can result in more clusters being identified, but also entail an increased risk of overfitting the data. The optimal trade-off between robustness of clustering and sensitivity was found by analysing the results of the algorithm with all possible values of the deepSplit parameter (that is, integer values from 0 to 4) on 100 subsamples of our data. In particular, in each subsample, we removed 10% of genes randomly selected before computing the dissimilarity matrix and applying the clustering algorithm.

The statistics of the Pearson gamma and the average silhouette width (computed with the 'cluster.stats' function included in the R package 'fpc', version 2.1-10)[43,44] of the subsamples (see Extended Data Fig. 3a,b) suggest that with 'deepSplit=2' a good compromise is reached between robustness and sensitivity for our data. We identified ten different clusters as well as two outlier cells that, although similar in gene expression to the mesodermal progenitor cells (cluster 4), were not assigned to any cluster by the algorithm, probably because of their relatively poor quality.

We then evaluated how specifically each gene is expressed in any given cluster. First, we found the differentially expressed genes (as described above) between all pairs of clusters. Marker genes for cluster *i* are expected to be significantly upregulated in *i* across all pairwise comparisons involving cluster *i*. The average rank of a marker gene across the pairwise comparisons provides a measure of how specifically the marker is expressed in the cluster. Extended Data Fig. 3c–f shows the expression values of marker genes for four different clusters. We provide the full list of markers in Supplementary Information Table 3. The clusters were visualized by using t-SNE (as implemented in the 'RtSNE' R package) on the dissimilarity matrix.

Single-cell trajectories in pseudospace: the anterior/posterior axis of the primitive streak. As discussed in the main text, cells allocated to cluster 4 (Fig. 1b–d) are cells that have probably exited the primitive streak only recently. We sought to align the cells along a pseudospatial trajectory representing the anterior–posterior axis of the primitive streak, which would allow us to identify the likely original locations of each cell along such an axis.

To do this we adopted an unsupervised approach: we did not use any prior information about marker genes, but selected the strongest signal present in this cluster of cells (controlling for potential batch effects) and later verified its biological meaning. We first used a diffusion map-based technique to reduce the dimensionality of the data set. Diffusion maps have recently been successfully applied to identify developmental trajectories in single-cell qPCR and RNA-seq data[14,15]. We used the implementation of the 'destiny' R package ('DiffusionMap' function) developed by Angerer *et al.*[45]. We restricted the analysis to genes that are highly variable among cells in the blue cluster and have an average expression above ten normalized read counts. The centred cosine similarity was used ('cosine' option in the 'DiffusionMap' function) and only the first two diffusion components (DC1 and DC2) were retained for downstream analysis.

In addition to biologically meaningful signals, batch effects (owing to cells being sorted and processed on different plates) can also be present and induce structure within the data. While in our data set the batch effect does not strongly influence the definition of different populations of cells, it might become relevant when finer structures within a single cluster of cells are considered (see Extended Data Fig. 6a). To tease apart the signals due to biological and batch effects, we computed the fraction of variance attributable to the batch effect along each direction in the diffusion space using a linear regression model. The direction 'orthogonal' to the batch effect, that is, the direction associated with the smallest fraction of variance explained by the batch effect, was considered as mostly driven by a biologically relevant signal. Hence, all cells were projected on this direction to obtain a 'pseudo-coordinate' representing the state of a cell relative to the biological process

captured by the diffusion map. The direction was identified by the angle α that it formed with the DC1 axis (Extended Data Fig. 6c).

Cells considered here are mostly from two batches including cells from the primitive streak stage (plate SLX-8408 and SLX-8409) and two batches including cells from the neural plate stage (plate SLX-8410 and SLX-8411; Extended Data Fig. 6b). For each of these two sets of batches, we computed the fraction of variance that can be explained by the batch covariate along any possible direction in the diffusion plot by using a linear regression model. The angles α_1 and α_2 corresponding to the directions orthogonal to the two batch effects are very close to each other (Extended Data Fig. 6c); we took the average value of α between these two angles to approximate the direction orthogonal to both batch effects.

Cells' coordinates in the diffusion space were projected along the direction identified by the average value of α, and this projection was interpreted as a 'pseudospace' coordinate representing the position of cells along the primitive streak (see main text and Fig. 3). We tested the robustness of such a pseudospace coordinate by repeating the same analysis with alternative dimensionality reduction techniques (t-SNE and independent component analysis), which gave highly correlated coordinates (see Extended Data Fig. 6d). A principal component analysis performed with a set of previously known markers for the anterior and posterior regions of the primitive streak also yielded a first component highly correlated with the pseudospace coordinate (see Extended Data Fig. 6h left panel). Moreover, the pseudospace coordinate had a positive (negative) correlation with the posterior (anterior) markers used (see Extended Data Fig. 6h right panel). These results strongly support the robustness of the signal we identified as well as its biological interpretation.

Once the pseudospace trajectory was defined, we selected genes that were differentially expressed along the trajectory. First, we removed all genes that were not detected in any cell. Then, for each gene, we fitted the $\log_{10}$(expression levels) (adding a pseudocount of 1) by using two local polynomial models: one with degree 0 and another with degree 2 ('locfit' function in 'locfit' R package, nearest neighbour component parameter equal to 1). The first, simpler model is better suited for genes that do not change their expression level along the trajectory. The second model has a greater number of parameters and is able to reproduce the more complex dynamics of genes that are differentially expressed.

We evaluated these two models by using the Akaike information criterion (AIC), a score that measures how well the data are reproduced by the model and includes a penalization for more complex models[46]. Better models according to this criterion correspond to smaller AIC scores.

To compute the AIC scores for the two models, we used the 'aic' function available in the 'locfit' R package, and then calculated the difference: $\Delta\mathrm{AIC} = \mathrm{AIC}(\mathrm{degree}=2) - \mathrm{AIC}(\mathrm{degree}=0)$. Negative values indicate that the more complex model with degree 2 local polynomials performs better, and therefore corresponds to genes that are more likely to be differentially expressed. Genes having a $\Delta\mathrm{AIC} < -2$ were considered to be significantly differentially expressed along the trajectory[46].

A hierarchical tree was built with the normalized expression patterns of the 462 differentially expressed genes (function 'hclust' with average linkage method and dissimilarity based on Spearman correlation) and a dynamic hybrid cut algorithm ('cutreeDynamic' function, minimum cluster size equal to 5) split this set of genes into three clusters according to the type of dynamics they have (see Fig. 3, Extended Data Fig. 6e and Supplementary Information Table 4).

Single-cell trajectories in pseudotime: the blood developmental trajectory. As discussed in the main text, clusters 7 and 8 (yellow and brown clusters in Fig. 1b, d) include blood progenitors at different stages of differentiation. By using a procedure analogous to the one described above, we aligned these cells along a trajectory representing embryonic blood development.

Extended Data Fig. 8a shows the diffusion plot with cells from the yellow and the brown clusters. Most of these cells come from plates SLX-8344 and SLX-8345 that were collected from embryos at neural plate and late head fold stages (see Extended Data Fig. 8b). With a linear regression model, where we controlled for biological parameters such as stage and sorting, we found the direction that correlates the least with the batch effect associated to these two plates and projected all cells onto it (Extended Data Fig. 8c). Note that the minimum correlation with the batch effect is achieved at a very small value of α ($\sim$10°, see Extended Data Fig. 8c), suggesting that the first diffusion component is mainly driven by a biologically meaningful signal and the batch effect plays a minor role here even at this more detailed scale of analysis. The new cell coordinate obtained from the projection was interpreted as a 'pseudotime' coordinate, which represents the differentiation stage of each cell along their journey towards erythroid fate. As expected, cells in the yellow cluster have a smaller pseudotime coordinate compared with the brown cluster that is mainly composed of more differentiated primitive erythroid cells. An analysis with alternative dimensionality reduction techniques yielded highly correlated pseudotime coordinates, suggesting the robustness of the signal (Extended Data Fig. 8d). Furthermore, our biological interpretation of the pseudotime coordinate is supported by the expression pattern of genes that are known to be upregulated or downregulated along the blood developmental trajectory, as is clear via principal component analysis (see Extended Data Fig. 8f).

By using the filtering and clustering procedure described in the previous section, we were able to detect 897 genes that were differentially expressed along the trajectory, which were divided in three clusters, each displaying a different type of dynamics (see Extended Data Fig. 8e and Supplementary Information Table 5).

Random Forest to allocate cells to previously identified clusters. Cells captured in the Tal1 experiment (testing data set) were allocated to the clusters we previously identified by using a Random Forest algorithm[47] (R package 'randomForest', version 4.6-12)[48] trained on the cells captured in the first experiment (training data set). The rank-normalized expression levels of all highly variable genes in the training data set were used as variables (the R function 'rank' was used for normalization, ties were averaged). The Random Forest algorithm was first used on the training data to assess variable importance with 1,000 classification trees. The 25% most important variables were selected to grow another set of 1,000 trees that were then used for the classification of the testing data set. With this filtered set of variables, the out-of-bag error estimate was $\sim$4.8%.

The quality of allocation of each cell in the testing data set was verified by computing the median of pairwise dissimilarities (defined as $(1-\rho)/2$, with ρ being the Spearman correlation) of that cell to all other cells in the training data allocated to the same cluster. Cells in the testing data set having a median pairwise dissimilarity larger than the maximum of the medians of pairwise dissimilarities of cells in the training data were considered to be 'unclassified' ($\sim$1.8% of all cells from the testing data set). For the identification of differentially expressed genes between clusters in the testing data, only cells that were confidently allocated to the clusters (that is, cells with a minimum difference of 10% probability between the best and the second best cluster allocation) were used.

Generation, maintenance and haematopoietic differentiation of Runx1–GFP/Gata1–mCherry ESCs. *Runx1*$^{GFP/+}$*Gata1*$^{mCherry/Y}$ ESCs were generated from morulae as described previously[49,50]. Cells were not tested for mycoplasma contamination. ESCs were grown on gelatinized plates (0.1% gelatin in water) at 37 °C and 5% CO_2 in ESC media (Knockout DMEM (Life Technologies) with 15% FCS (batch-tested for ESC culture; Life Technologies), 2 mM L-glutamine (PAA Laboratories), 0.5% P/S, 0.1 mM β-mercaptoethanol (Life Technologies) and 10^3 U/ml recombinant LIF (ORF Genetics)). Cells were passaged with TrypLE Express dissociation reagent (Life Technologies) every 1–3 days. ESCs were differentiated as embryoid bodies as previously described[31,51]. Embryoid bodies were harvested into Falcon tubes after 5 days of culture and dissociated with TrypLE Express dissociation reagent and prepared for FACS.

ChIP-seq. ChIP was performed as described[52] with modifications for low cell numbers[53]. Approximately 7×10^6 FACS-sorted day 5 embryoid body cells (Runx1-ires–GFP$^+$/Gata1–mCherry$^+$; Extended Data Fig. 9a) per ChIP were cross-linked using formaldehyde to a final concentration of 1%. As samples were pooled from several sorts, isolated nuclei were frozen on dry ice-cold isopropanol and stored at −80 °C. During the immunoprecipitation step, 4 μl recombinant histone 2B (New England Biolabs) and 1 μl of mouse RNA (Qiagen; diluted 1/5 in IP dilution buffer) were added as carriers, followed by 7 μg of primary antibody (rabbit anti-Gata1, Abcam ab11963). Sequencing libraries were prepared using the TruSeq Kit (Illumina) for high throughput sequencing on an Illumina HiSeq 2500, according to the manufacturer's instructions, with size selection for fragments of 150–400 bp.

ChIP-seq mapping and analysis. Alignment of the ChIP-seq reads to the mouse mm10 genome, quality control and peak calling were performed according to the data pipeline set out by Sanchez-Castillo *et al.*[54]. Peak calling was performed using MACS2[55] with $P = 1 \times 10^{-6}$. Post-processing using in-house scripts converted the peak coordinates to 400 bp on the basis of peak summits given in the MACS output. Coordinates of genomic regions that lie at the end of chromosomes and/or in repeat regions were discarded from the final high-confidence peak lists. PolyAPeak[56] was run in R to remove abnormally shaped peaks. Peaks were assigned to genes using an in-house script according to whether they overlapped with a known TSS or fell within 50 kb each side of a gene.

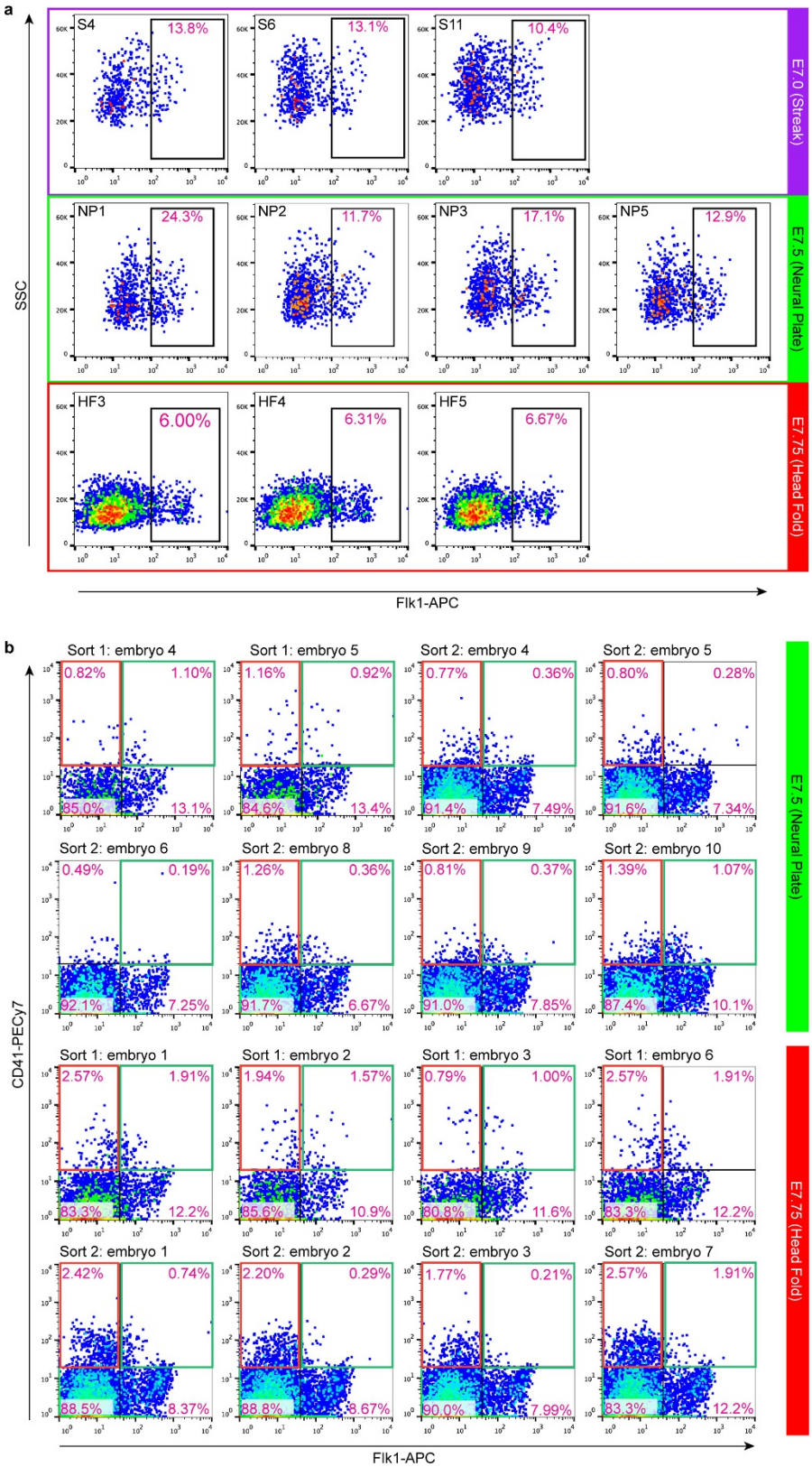
***In situ* hybridization.** Whole-mount *in situ* hybridization for *Tal1* was performed as described previously[57]. An *in situ* hybridization probe for *Tal1* was synthesized using published sequence (*Tal1* 860-1428, accession number M59764) with the DIG RNA labelling kit (Roche).

Code availability. All data were analysed with standard programs and packages, as detailed above. Code is available on request.

30. Downs, K. M. & Davies, T. Staging of gastrulating mouse embryos by morphological landmarks in the dissecting microscope. *Development* **118,** 1255–1266 (1993).
31. 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).

32. Elefanty, A. G. *et al.* Characterization of hematopoietic progenitor cells that express the transcription factor SCL, using a lacZ "knock-in" strategy. *Proc. Natl Acad. Sci. USA* **95,** 11897–11902 (1998).
33. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* **9,** 171–181 (2014).
34. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).
35. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Rev. Genet.* **16,** 133–145 (2015).
36. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).
37. van der Maaten, L. Barnes-Hut-SNE. Preprint at http://arxiv.org/pdf/1301.3342 (2013).
38. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
39. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* **10,** 1093–1095 (2013).
40. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnol.* **33,** 155–160 (2015).
41. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
42. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24,** 719–720 (2008).
43. Halkidi, M., Batistakis, Y. & Vazirgiannis, M. On clustering validation techniques. *J. Intell. Inf. Syst.* **17,** 107–145 (2001).
44. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20,** 53–65 (1987).
45. Angerer, P. *et al.* destiny – diffusion maps for large-scale single-cell data in R. *bioRxiv* http://dx.doi.org/10.1101/023309 (2015).
46. Burnham, K. P. & Anderson, D. R. in *Model Selection and Multimodel Inference A Practical Information-Theoretic Approach* 2nd edn, Ch. 2 (Springer, 2002).
47. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).
48. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2,** 18–22 (2002).
49. Bryja, V., Bonilla, S. & Arenas, E. Derivation of mouse embryonic stem cells. *Nature Protocols* **1,** 2082–2087 (2006).
50. Tanaka, Y. *et al.* Circulation-independent differentiation pathway from extraembryonic mesoderm toward hematopoietic stem cells via hemogenic angioblasts. *Cell Reports* **8,** 31–39 (2014).
51. Sroczynska, P., Lancrin, C., Pearson, S., Kouskoff, V. & Lacaud, G. In vitro differentiation of mouse embryonic stem cells as a model of early hematopoietic development. *Methods Mol. Biol.* **538,** 317–334 (2009).
52. Wilson, N. K. *et al.* Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7,** 532–544 (2010).
53. Zwart, W. *et al.* A carrier-assisted ChIP-seq method for estrogen receptor-chromatin interactions from breast cancer core needle biopsy samples. *BMC Genomics* **14,** 232 (2013).
54. Sánchez-Castillo, M. *et al.* CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res.* **43,** D1117–D1123 (2015).
55. Zhang, Y. *et al.* Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9,** R137 (2008).
56. Wu, H. & Ji, H. PolyaPeak: detecting transcription factor binding sites from ChIP-seq using peak shape information. *PLoS ONE* **9,** e89694 (2014).
57. Wilkinson, D. G. *In Situ Hybridization* (Oxford Univ. Press, 1999).
58. Beddington, R. S. & Robertson, E. J. Axis development and early asymmetry in mammals. *Cell* **96,** 195–209 (1999).
59. Du, J. *et al. O*-fucosylation of thrombospondin type 1 repeats restricts epithelial to mesenchymal transition (EMT) and maintains epiblast pluripotency during mouse gastrulation. *Dev. Biol.* **346,** 25–38 (2010).
60. Donnison, M. *et al.* Loss of the extraembryonic ectoderm in *Elf5* mutants leads to defects in embryonic patterning. *Development* **132,** 2299–2308 (2005).
61. Mitsunaga, K. *et al.* Loss of PGC-specific expression of the orphan nuclear receptor ERR-β results in reduction of germ cell number in mouse embryos. *Mech. Dev.* **121,** 237–246 (2004).
62. Baldwin, H. S. *et al.* Platelet endothelial cell adhesion molecule-1 (PECAM-1/CD31): alternatively spliced, functionally distinct isoforms expressed during mammalian cardiovascular development. *Development* **120,** 2539–2553 (1994).
63. Naiche, L. A., Arora, R., Kania, A., Lewandoski, M. & Papaioannou, V. E. Identity and fate of *Tbx4*-expressing cells reveal developmental cell fate decisions in the allantois, limb, and external genitalia. *Dev. Dyn.* **240,** 2290–2300 (2011).
64. Tamplin, O. J. *et al.* Microarray analysis of *Foxa2* mutant mouse embryos reveals novel gene expression and inductive roles for the gastrula organizer and its derivatives. *BMC Genomics* **9,** 511 (2008).
65. Vincent, S. D. *et al. Prdm1* functions in the mesoderm of the second heart field, where it interacts genetically with *Tbx1*, during outflow tract morphogenesis in the mouse embryo. *Hum. Mol. Genet.* **23,** 5087–5101 (2014).
66. Morkel, M. *et al.* β-Catenin regulates Cripto- and Wnt3-dependent gene expression programs in mouse axis and mesoderm formation. *Development* **130,** 6283–6294 (2003).
67. Niwa, H., Miyazaki, J. & Smith, A. G. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nature Genet.* **24,** 372–376 (2000).
68. Pearce, J. J. H. & Evans, M. J. Mml, a mouse Mix-like gene expressed in the primitive streak. *Mech. Dev.* **87,** 189–192 (1999).
69. Pennisi, D. *et al.* Mutations in Sox18 underlie cardiovascular and hair follicle defects in ragged mice. *Nature Genet.* **24,** 434–437 (2000).
70. Gordon, E. J., Gale, N. W. & Harvey, N. L. Expression of the hyaluronan receptor LYVE-1 is not restricted to the lymphatic vasculature; LYVE-1 is also expressed on embryonic blood vessels. *Dev. Dyn.* **237,** 1901–1909 (2008).
71. Kallianpur, A. R., Jordan, J. E. & Brandt, S. J. The SCL/TAL-1 gene is expressed in progenitors of both the hematopoietic and vascular systems during embryogenesis. *Blood* **83,** 1200–1208 (1994).
72. 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).
73. Tanaka, Y. *et al.* The transcriptional programme controlled by Runx1 during early embryonic blood development. *Dev. Biol.* **366,** 404–419 (2012).
74. North, T. *et al.* Cbfa2 is required for the formation of intra-aortic hematopoietic clusters. *Development* **126,** 2563–2575 (1999).
75. Palis, J., McGrath, K. E. & Kingsley, P. D. Initiation of hematopoiesis and vasculogenesis in murine yolk sac explants. *Blood* **86,** 156–163 (1995).
76. Drissen, R. *et al.* The erythroid phenotype of EKLF-null mice: defects in hemoglobin metabolism and membrane stability. *Mol. Cell. Biol.* **25,** 5205–5214 (2005).
77. Southwood, C. M., Downs, K. M. & Bieker, J. J. Erythroid Krüppel-like factor exhibits an early and sequentially localized pattern of expression during mammalian erythroid ontogeny. *Dev. Dyn.* **206,** 248–259 (1996).
78. Silver, L. & Palis, J. Initiation of murine embryonic erythropoiesis: a spatial analysis. *Blood* **89,** 1154–1164 (1997).
79. Lanctôt, C., Lamolet, B. & Drouin, J. The bicoid-related homeoprotein Ptx1 defines the most anterior domain of the embryo and differentiates posterior from anterior lateral mesoderm. *Development* **124,** 2807–2817 (1997).
80. Lania, G., Ferrentino, R. & Baldini, A. TBX1 represses Vegfr2 gene expression and enhances the cardiac fate of VEGFR2+ cells. *PLoS ONE* **10,** e0138525 (2015).
81. Brown, C. B. *et al.* Cre-mediated excision of *Fgf8* in the *Tbx1* expression domain reveals a critical role for *Fgf8* in cardiovascular development in the mouse. *Dev. Biol.* **267,** 190–202 (2004).
82. Brennan, J. *et al.* Nodal signalling in the epiblast patterns the early mouse embryo. *Nature* **411,** 965–969 (2001).
83. Meno, C. *et al.* Mouse Lefty2 and zebrafish antivin are feedback inhibitors of nodal signaling during vertebrate gastrulation. *Mol. Cell* **4,** 287–298 (1999).
84. Bessho, Y. *et al.* Dynamic expression and essential functions of Hes7 in somite segmentation. *Genes Dev.* **15,** 2642–2647 (2001).
85. Oginuma, M., Niwa, Y., Chapman, D. L. & Saga, Y. Mesp2 and Tbx6 cooperatively create periodic patterns coupled with the clock machinery during mouse somitogenesis. *Development* **135,** 2555–2562 (2008).
86. Forlani, S., Lawson, K. A. & Deschamps, J. Acquisition of Hox codes during gastrulation and axial elongation in the mouse embryo. *Development* **130,** 3807–3819 (2003).
87. Zeigler, B. M. *et al.* The allantois and chorion, when isolated before circulation or chorio-allantoic fusion, have hematopoietic potential. *Development* **133,** 4183–4192 (2006).
88. Downs, K. M., Hellman, E. R., McHugh, J., Barrickman, K. & Inman, K. E. Investigation into a role for the primitive streak in development of the murine allantois. *Development* **131,** 37–55 (2004).
89. Caprioli, A., Jaffredo, T., Gautier, R., Dubourg, C. & Dieterlen-Lièvre, F. Blood-borne seeding by hematopoietic and endothelial precursors from the allantois. *Proc. Natl Acad. Sci. USA* **95,** 1641–1646 (1998).
90. van Nes, J. *et al.* The *Cdx4* mutation affects axial development and reveals an essential role of Cdx genes in the ontogenesis of the placental labyrinth in mice. *Development* **133,** 419–428 (2006).
91. Yang, J. T., Rayburn, H. & Hynes, R. O. Cell adhesion events mediated by α_4 integrins are essential in placental and cardiac development. *Development* **121,** 549–560 (1995).
92. Solloway, M. J. & Robertson, E. J. Early embryonic lethality in *Bmp5;Bmp7* double mutant mice suggests functional redundancy within the 60A subgroup. *Development* **126,** 1753–1768 (1999).
93. Drake, C. J. & Fleming, P. A. Vasculogenesis in the day 6.5 to 9.5 mouse embryo. *Blood* **95,** 1671–1679 (2000).
94. Lee, D. *et al.* ER71 acts downstream of BMP, Notch, and Wnt signaling in blood and vessel progenitor specification. *Cell Stem Cell* **2,** 497–507 (2008).
95. Carapuço, M., Nóvoa, A., Bobola, N. & Mallo, M. *Hox* genes specify vertebral types in the presomitic mesoderm. *Genes Dev.* **19,** 2116–2121 (2005).
96. Zhang, H. *et al.* Expression of podocalyxin separates the hematopoietic and vascular potentials of mouse embryonic stem cell-derived mesoderm. *Stem Cells* **32,** 191–203 (2014).

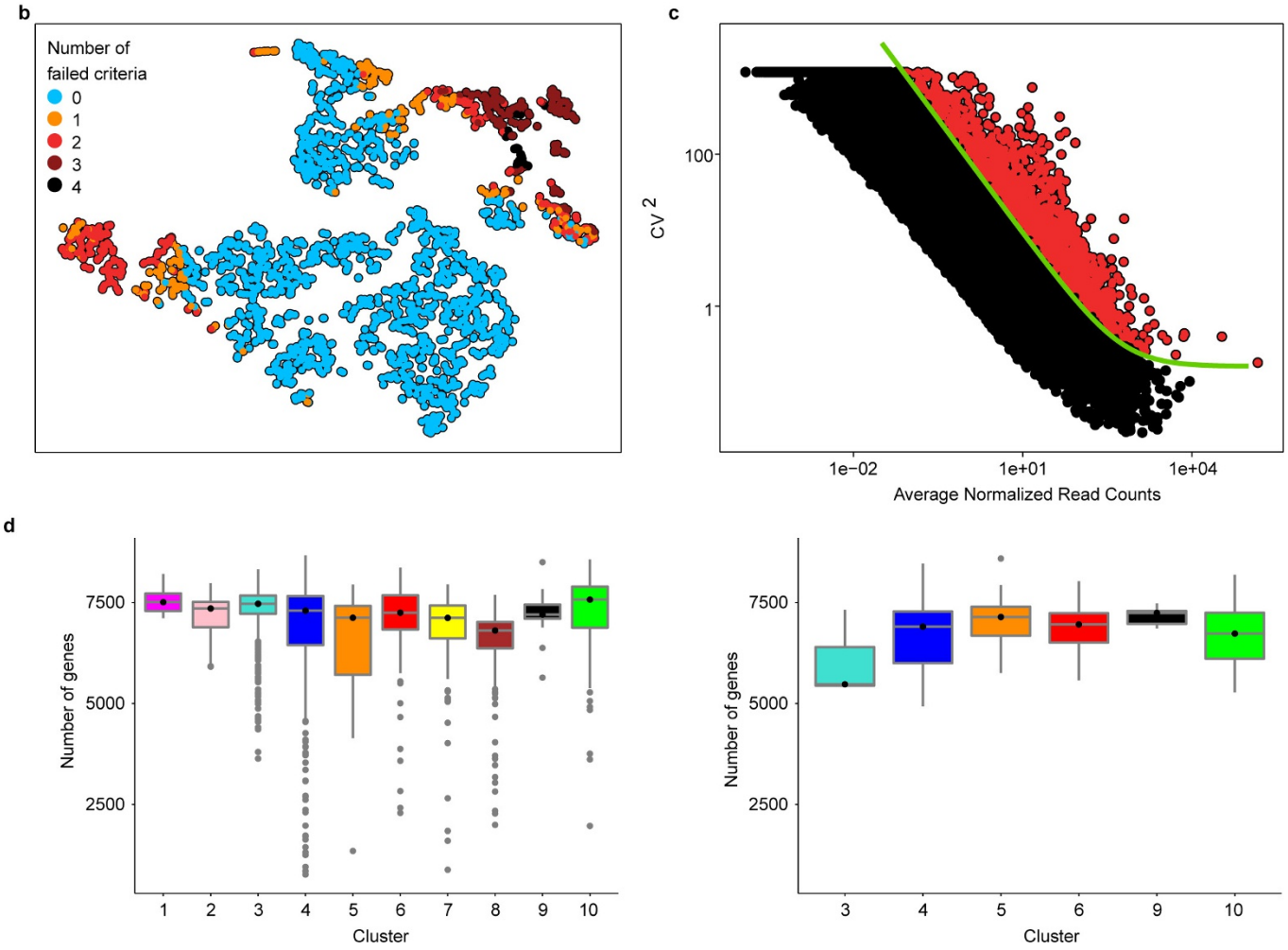
97. Herrmann, B. G. Expression pattern of the *Brachyury* gene in whole-mount TWis/TWis mutant embryos. *Development* **113,** 913–917 (1991).
98. Weidgang, C. E. *et al.* TBX3 directs cell-fate decision toward mesendoderm. *Stem Cell Rep.* **1,** 248–265 (2013).
99. Perea-Gómez, A., Shawlot, W., Sasaki, H., Behringer, R. R. & Ang, S. *HNF3β* and *Lim1* interact in the visceral endoderm to regulate primitive streak formation and anterior-posterior polarity in the mouse embryo. *Development* **126,** 4499–4511 (1999).
100. Saga, Y. *et al.* MesP1 is expressed in the heart precursor cells and required for the formation of a single heart tube. *Development* **126,** 3437–3447 (1999).
101. Trimborn, T., Gribnau, J., Grosveld, F. & Fraser, P. Mechanisms of developmental control of transcription in the murine α- and β-globin loci. *Genes Dev.* **13,** 112–124 (1999).
102. Kingsley, P. D., Malik, J., Fantauzzo, K. A. & Palis, J. Yolk sac-derived primitive erythroblasts enucleate during mammalian embryogenesis. *Blood* **104,** 19–25 (2004).
103. Hodge, D. *et al.* A global role for EKLF in definitive and primitive erythropoiesis. *Blood* **107,** 3359–3370 (2006).
104. Isern, J. *et al.* Single-lineage transcriptome analysis reveals key regulatory pathways in primitive erythroid progenitors in the mouse embryo. *Blood* **117,** 4924–4934 (2011).
105. Joshi, A., Hannah, R., Diamanti, E. & Göttgens, B. Gene set control analysis predicts hematopoietic control mechanisms from genome-wide transcription factor binding data. *Exp. Hematol.* **41,** 354–366.e14 (2013).
106. Goode, D. K. *et al.* Dynamic gene regulatory networks drive hematopoietic specification and differentiation. *Dev. Cell* **36,** 572–587 (2016).



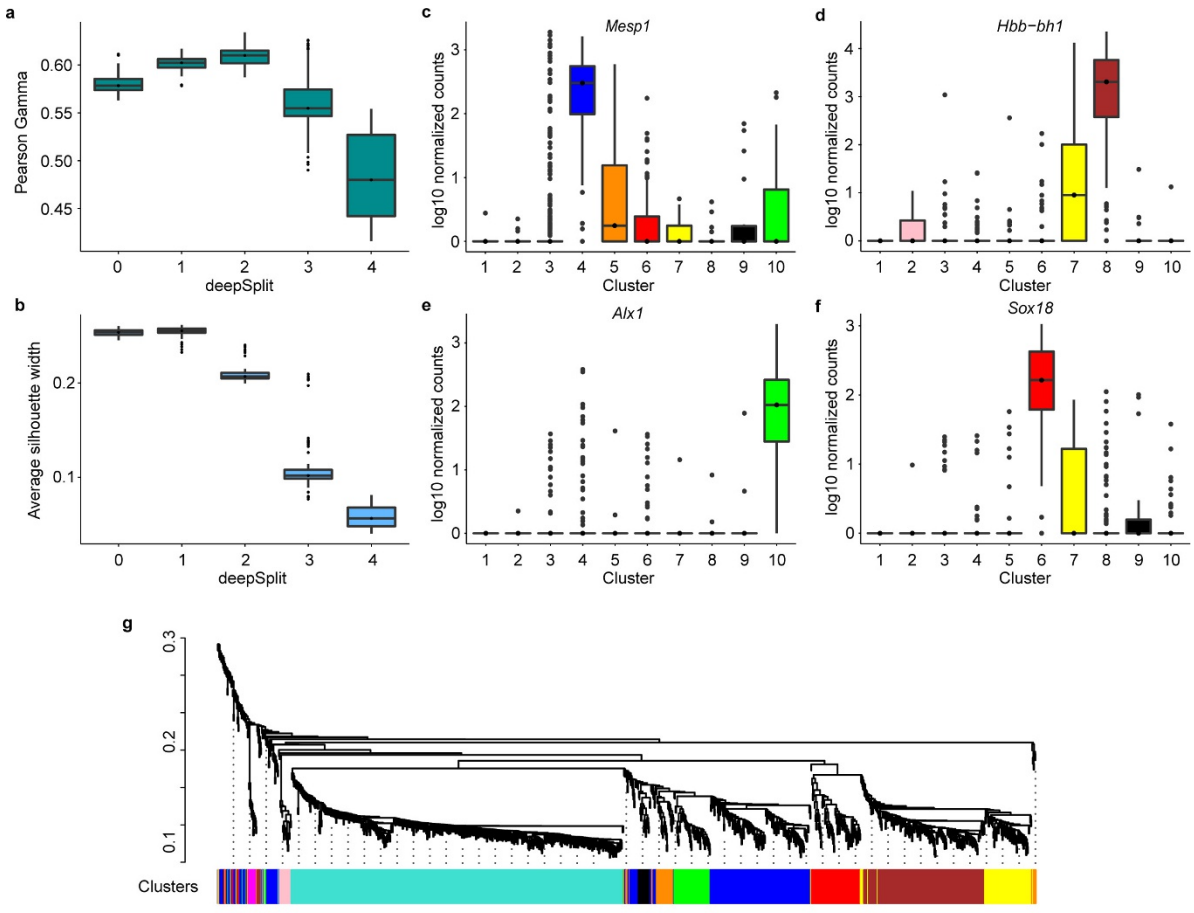
Extended Data Figure 1 | FACS of single cells. a, Flk1$^+$ cells were sorted from three embryos at primitive streak and head fold stages and four embryos at neural plate stage. Labels such as 'S4' refer to the embryo number in the metadata available online at http://gastrulation.stemcells.cam.ac.uk/scialdone2016. **b**, CD41$^+$Flk1$^-$ cells (red gate) and CD41$^+$Flk1$^+$ (green gate) cells were sorted from eight embryos each at neural plate and head fold stages (see Fig. 1 for cell numbers). Each stage was sorted on two occasions. Labels above FACS plots refer to the sort and embryo number in the metadata available online, as above. In all plots, pink text indicates the percentage of cells in that gate.

a

		E6.5	S	NP	HF
	Total cells	600*	13182†	20299†	29805†
Epiblast	number sorted	501	ND	ND	ND
Flk1	estimated number	ND	195†	241†	704†
	number sorted	ND	138	159	139
CD41+Flk1+	percentage	ND	ND	0.96	1.81
	estimated number	ND	ND	538	538
	number sorted	ND	ND	45	90
CD41+	percentage	ND	ND	0.66	0.79
	estimated number	ND	ND	134	235
	number sorted	ND	ND	55	78



Extended Data Figure 2 | Quality control of single-cell RNA-seq data. **a**, Table showing numbers and estimates of numbers of cells of different phenotypes present in embryos between E6.5 and E7.75 (head fold stage) and numbers sorted for this study. *Total cell numbers for E6.5 are from Beddington and Robertson (1999)[58]. †Total numbers and numbers of Flk1^{+} cells are from Moignard *et al.*, (2015)[15]. Percentages of cells expressing Flk1 and/or CD41 at neural plate and head fold stages are the average values from the embryos used in this study and were used to calculate the estimated numbers present in embryos from the total cell numbers. ND, not done. **b**, t-SNE representation of the five metrics used to assess the quality of all 2,208 sorted cells from the wild-type and Tal1 experiments. Only cells that passed all criteria (blue circles) were used for downstream analysis. **c**, Squared coefficient of variation (CV^2) as a function of the mean normalized counts (μ) for genes across all cells. The green line shows the fit $CV^2 = a_1/\mu + \alpha_0$. All highly variable genes (with an adjusted P value < 0.1) are marked by red circles. **d**, Number of genes detected (that is, with more than ten normalized read counts) in cells across the different clusters in the WT (left) and the *Tal1*$^{-/-}$ (right) mice. Boxes indicate the median and interquartile range.

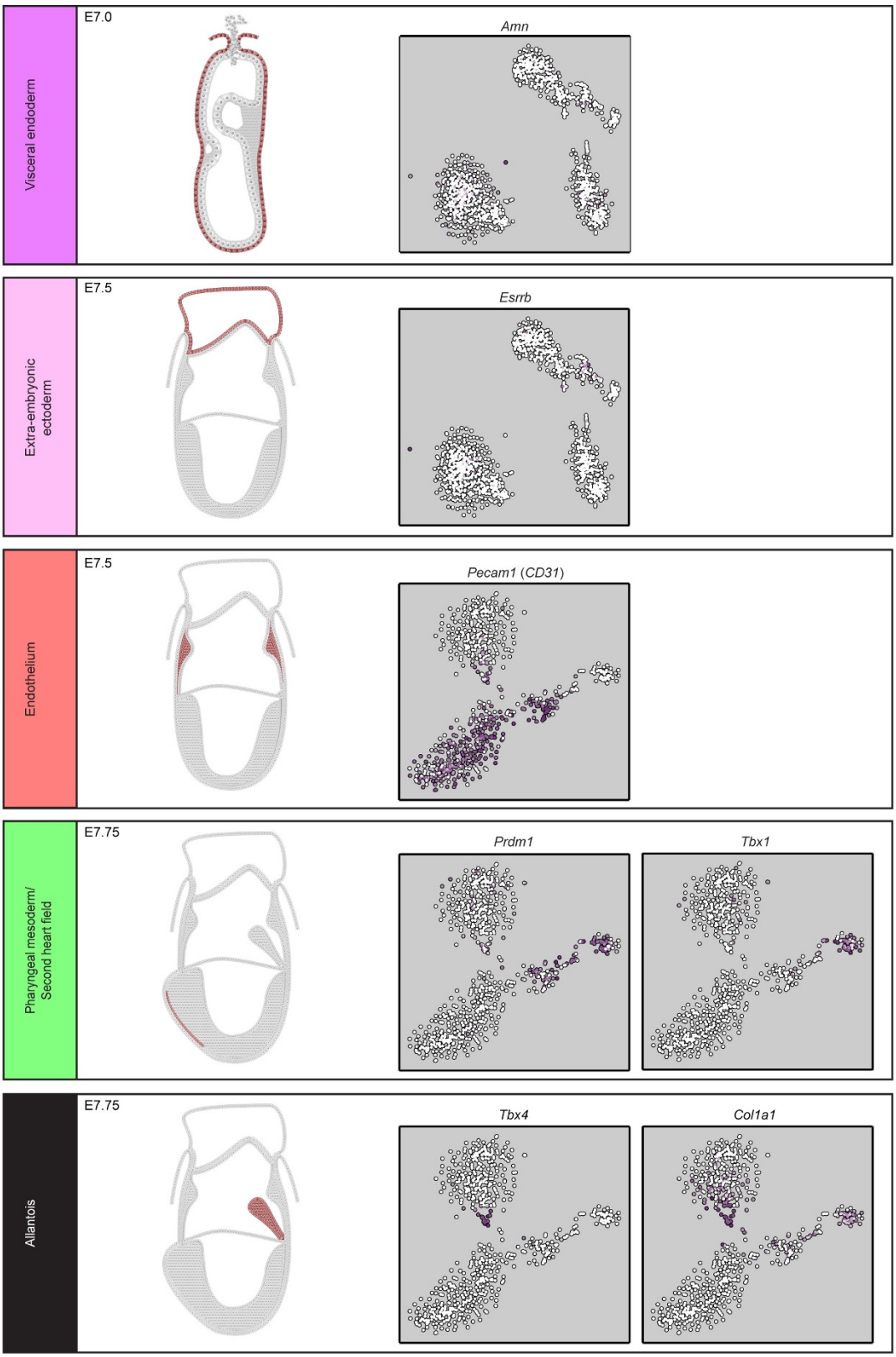


h

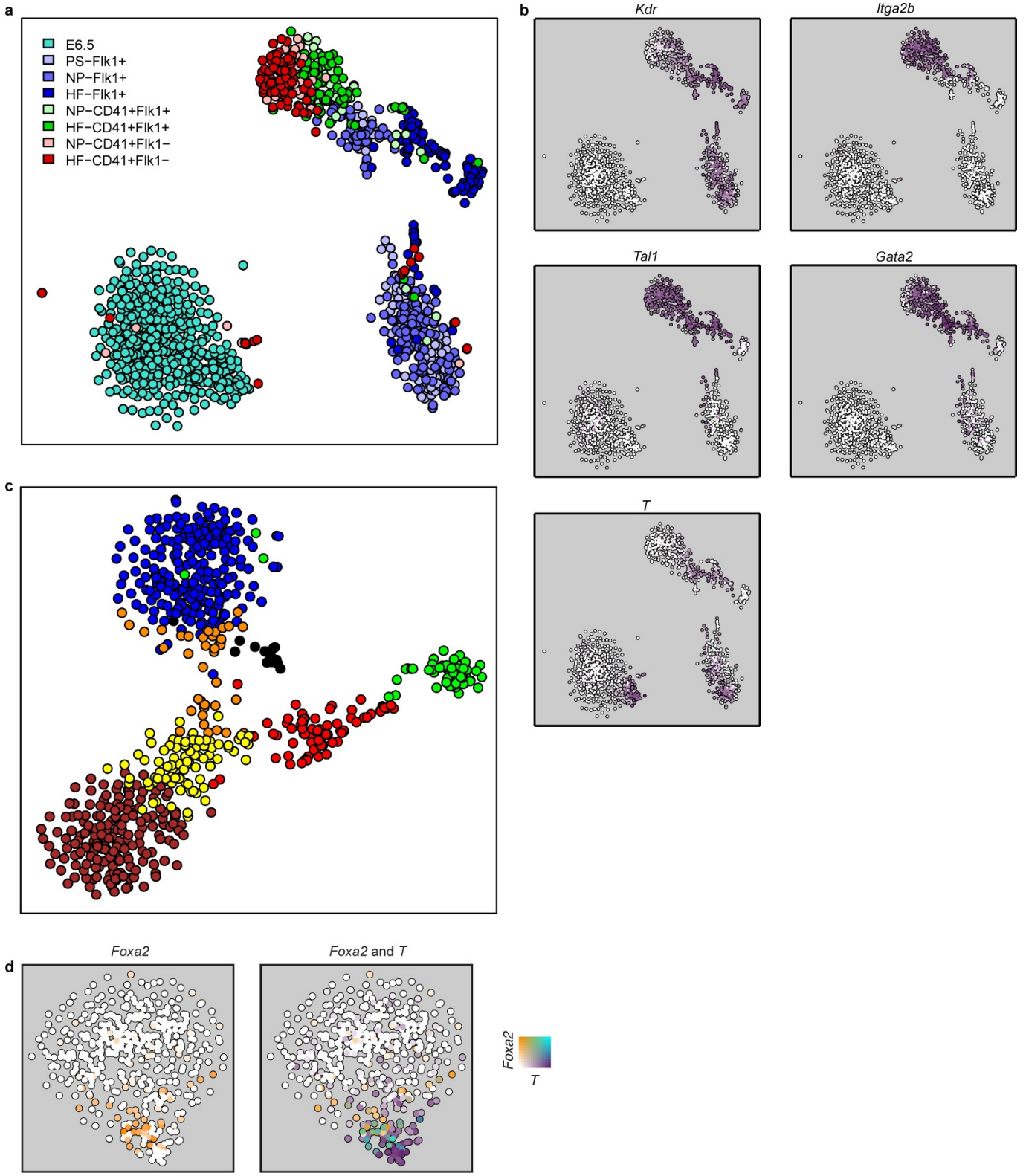
Cluster number	Number of cells	Cluster colour	Main stage	Assigned ID	Key genes and evidence
1	12	Magenta	E6.5	Visceral endoderm	*Amn* and *Cubn* [59]
2	17	Pink	E6.5	Extra-embryonic ectoderm	*Essrb* [61] *Elf5* [60]
3	492	Turquoise	E6.5	Epiblast and primitive streak	*Nanog* [66] *Pou5f1 (Oct4)* [67]
4	216	Blue	S/NP	Mesodermal progenitors	*Mesp1* [100] *Mixl1* [68] *Wnt3* [59]
5	40	Orange	NP	Posterior mesoderm	
6	78	Red	HF	Endothelium (developing YS vasculature)	*Sox18* [69] *Pecam1* [62] *Lyve1* [20,70]
7	84	Yellow	NP	Blood progenitors	*Tal1* [71,72] *Runx1* [50,73,74]
8	187	Brown	HF	Blood cells (primitive erythroid)	*Hbb-bh1* [75] *Klf1* [76,77] *Gata1* [78]
9	18	Black	HF	Allantois	*Pitx1* [79] *Tbx4* [63] *Col1a1* [64]
10	59	Green	HF	Pharyngeal mesoderm	*Tbx1* [80] *Prdm1* [65]

Extended Data Figure 3 | See next page for caption.

Extended Data Figure 3 | Identifying cell clusters. The dynamic hybrid cut algorithm was used with all possible values of the 'deepSplit' parameter on 100 bootstrapped subsamples. **a**, **b**, To assess the quality of the clustering, the Pearson gamma (**a**) and the average silhouette width (**b**) were calculated. Higher values of these parameters correspond to better clustering. The Pearson gamma represents the correlation between the dissimilarity of samples and a binary variable that equals 0 for pairs of samples in the same cluster and 1 for samples in different clusters. The average silhouette width measures the average separation between neighbouring clusters[43,44]. At 'deepSplit' = 2 the Pearson gamma is highest whereas the average silhouette width begins to decrease. This suggests that at such a value of the 'deepSplit' parameter a good compromise between robustness and sensitivity is achieved. The Pearson gamma and the average silhouette width were computed with the R function 'cluster.stats' in the 'fpc' package (version 2.1-9). **c**–**f**, Examples of marker genes for four clusters: *Mesp1* for cluster 4 (top-ranked marker) (**c**), *Hbb-bh1* for cluster 8 (fourth-ranked) (**d**), *Alx1* for cluster 10 (top-ranked) (**e**) and *Sox18* for cluster 6 (second-ranked) (**f**). The *y* axis shows the $\log_{10}$-normalized expression of the genes. For **a**–**f**, boxes indicate the median and interquartile range. **g**, Dendrogram showing the clustering of the cells in the first experiment. The colours at the bottom indicate the cluster each cell was assigned to by the dynamic hybrid cut algorithm. Cluster assignment was used to sort cells in Fig. 1b. **h**, Identities were assigned to the ten clusters in Fig. 1c on the basis of the expression of key genes[20,50,59–80] associated with various mesodermal lineages or spatial locations within the embryo.



Extended Data Figure 4 | Expression of key marker genes in E7.0–7.75 embryos. Schematic representations of expression patterns were generated from published *in situ* hybridization data (see citations) for key markers of clusters 1 (magenta, visceral endoderm[59]), 2 (pink, extra-embryonic ectoderm[61]), 6 (red, yolk sac endothelium[62]), 9 (black, allantois[63,64]) and 10 (green, second heart field[65,81]). Anterior is shown on the left and posterior on the right. Also shown is the t-SNE for all 1,205 cells or 682 cells from E7.0 onwards (primitive streak, neural plate and head fold stages) indicating expression of each gene (white, low; purple, high).



Extended Data Figure 5 | Expression of key genes used for sorting single cells. a, t-SNE as in Fig. 1 showing the sorting strategy for each of the 1,205 cells. **b**, Expression of Flk1 (*Kdr*), CD41 (*Itga2b*), Scl (*Tal1*), *Gata2* and *T* (Brachyury) superimposed onto the t-SNE. **c**, t-SNE showing only the 682 cells from primitive streak, neural plate and head fold stages, coloured according to cluster as in Fig. 1c, e. **d**, t-SNE for the 481 E6.5 cells in cluster 3, as in Fig. 2a. Each point is coloured by expression of *T* and *Foxa2*.

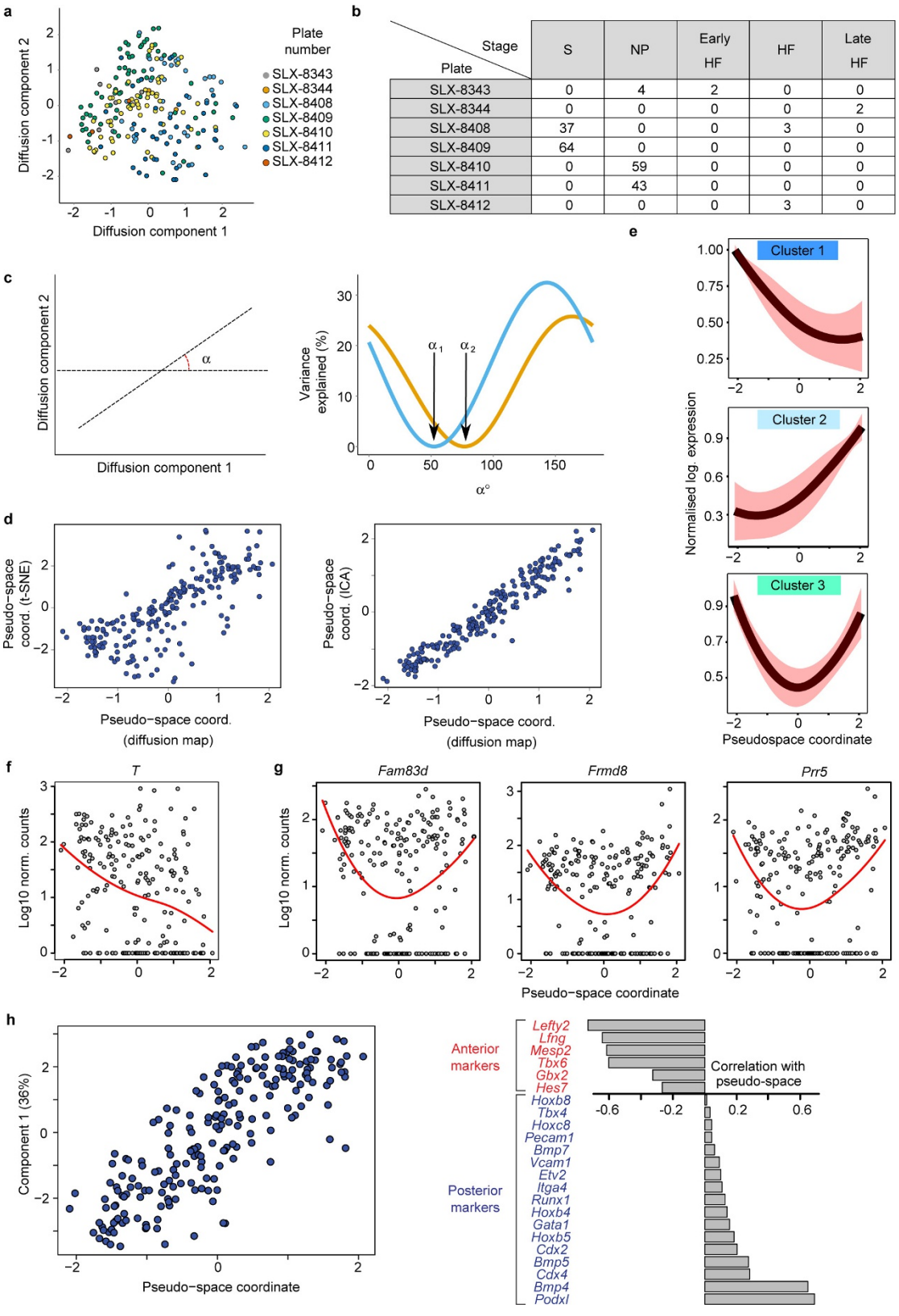
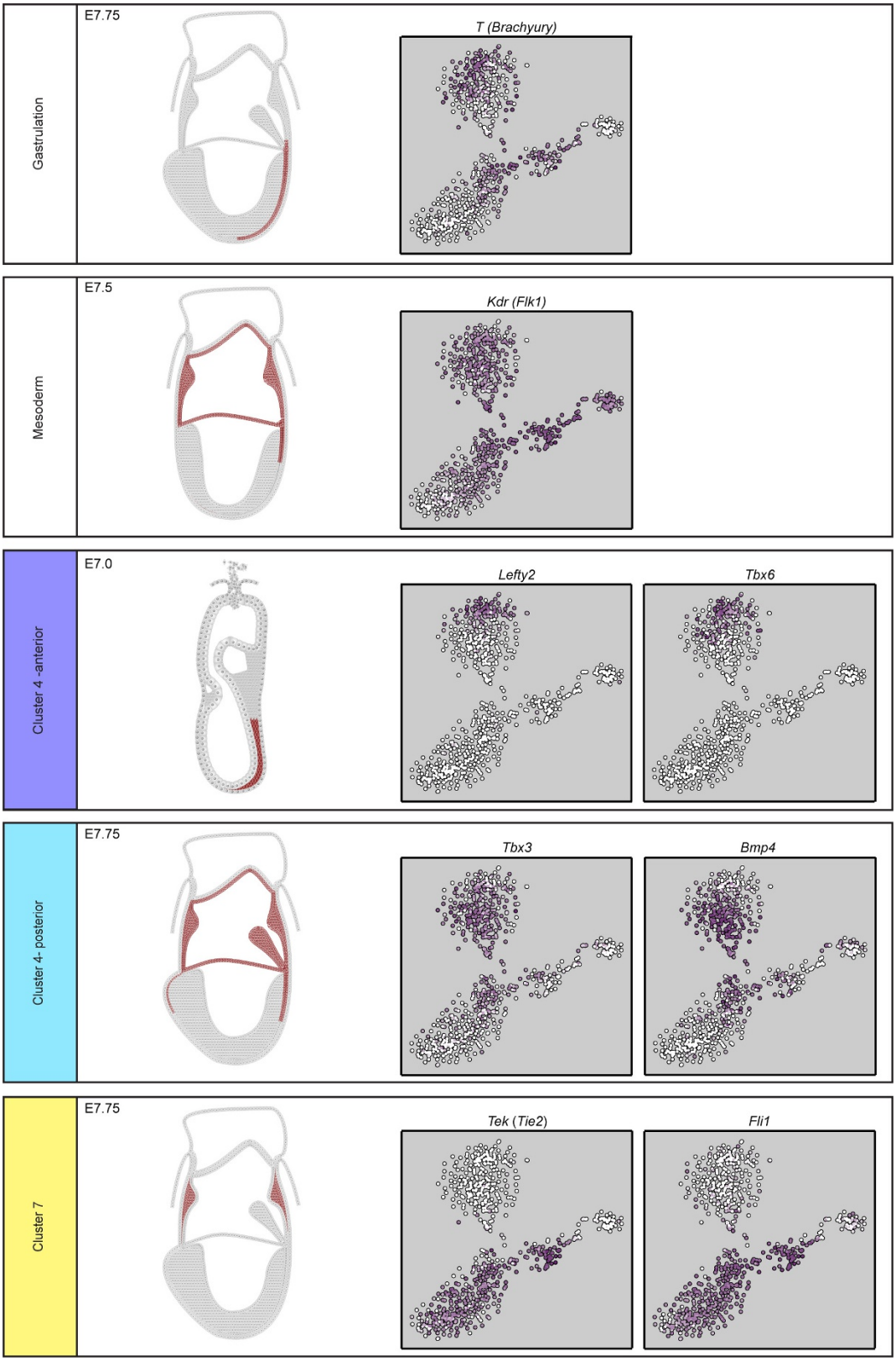


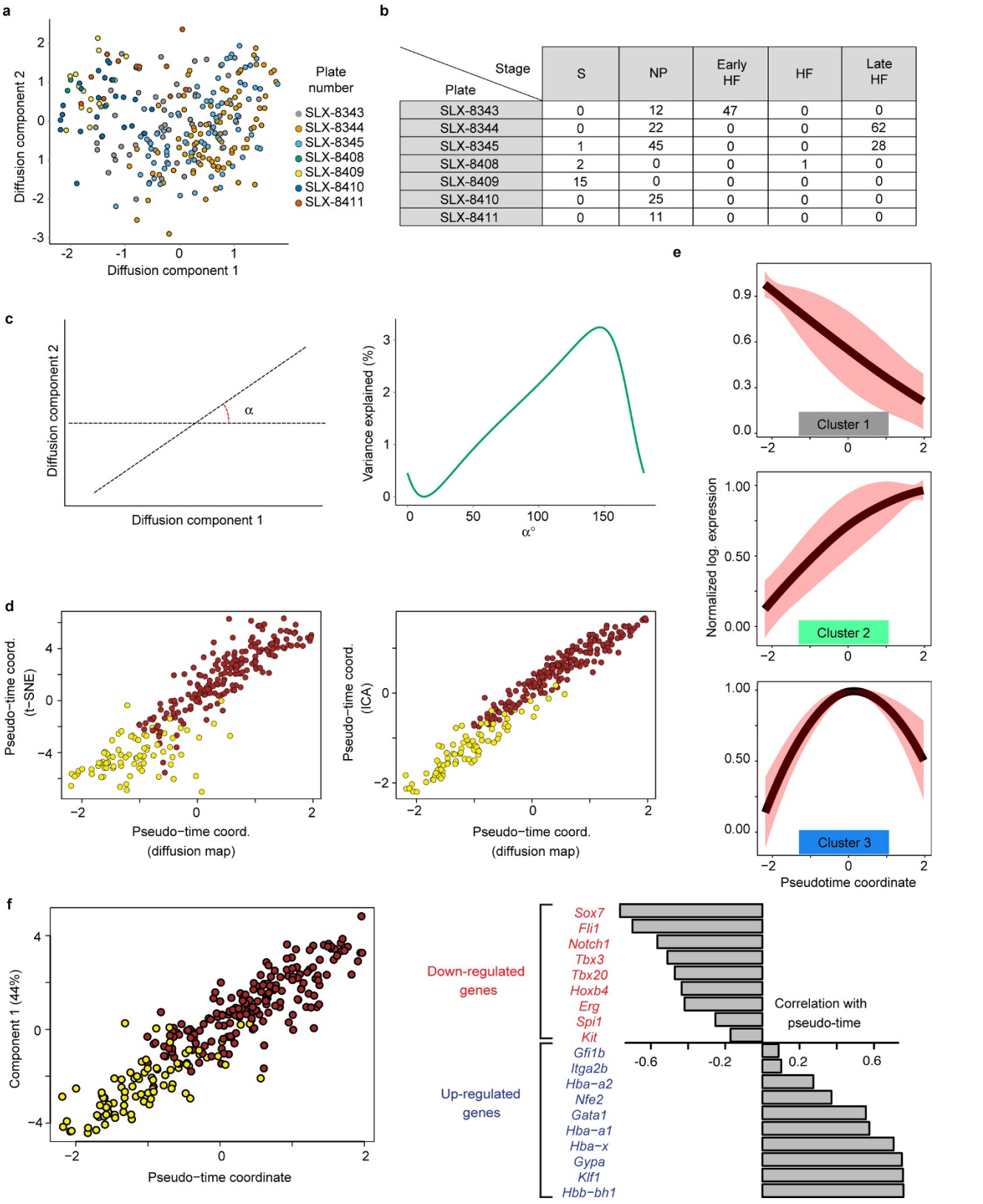
Plate \ Stage	S	NP	Early HF	HF	Late HF
SLX-8343	0	4	2	0	0
SLX-8344	0	0	0	0	2
SLX-8408	37	0	0	3	0
SLX-8409	64	0	0	0	0
SLX-8410	0	59	0	0	0
SLX-8411	0	43	0	0	0
SLX-8412	0	0	0	3	0

Extended Data Figure 6 | See next page for caption.

Extended Data Figure 6 | Pseudospace analysis of cluster 4 correlates with anterior–posterior position along the primitive streak. a, Diffusion plot of the 216 cells in cluster 4. Different colours correspond to different plates and different lanes of flow cells. **b**, Table showing the number of cells in each stage analysed on the different lanes of flow cells (S, primitive streak; NP, neural plate; HF, head fold). **c**, A direction in the diffusion space can be identified by the angle α that it forms with the first diffusion component (left panel). For each value of α the right panel shows the percentage of variance explained by the batch effect associated to plates SLX-8408 and SLX-8409 (orange line) and plates SLX-8410 and SLX-8411 (blue line). Labels α_1 and α_2 are the angles corresponding to directions that correlate the least with the batch effect (that is, variance explained by the batch effect is minimum). **d**, The use of alternative dimensionality reduction techniques results in the identification of highly correlated pseudospace coordinates. A t-SNE projection of the dissimilarity matrix was performed (perplexity set to 50), and the direction corresponding to the pseudospace coordinate was estimated by minimizing the correlation with the batch effect (left panel; Spearman correlation between the two pseudospace coordinates 0.79, $P < 2.2 \times 10^{-16}$). Independent component analysis was performed on the dissimilarity matrix with the 'fastICA' R function, and three independent components (corresponding to the two batch effects and the biological effect) were estimated. The presumptive pseudospace coordinate is the component having the smallest correlation with the batch effects (right panel; Spearman correlation coefficient is 0.97, $P < 2.2 \times 10^{-16}$). **e**, Plots showing the average expression of genes in clusters 1–3 of Fig. 3c along the pseudospace axis. Gene expression levels are normalized between 0 and 1. Dark red lines indicate the normalized mean expression levels of genes in each cluster as obtained from the fitting procedure and red shaded area indicates standard deviation. **f**, Expression of *T* as function of the pseudospace coordinate. **g**, Gene expression levels for example genes showing high-low-high expression pattern across the blue cluster. In **f** and **g**, putative anterior cells are to the left and posterior to the right. Each dot represents a cell and red lines indicate fits based on local polynomial functions (see Methods). **h**, We performed principal component analysis on the cells in cluster 4 by using markers of pre-somitic mesoderm as anterior mesoderm markers and genes expressed in haemato-vascular and allantoic mesoderm as posterior markers[82–95], as well as *Podxl* which was shown to separate distinct Flk1$^+$ mesodermal lineages[96]. The first component explained 36% of the total variance and was highly correlated with the pseudospace coordinate (left; Spearman rank correlation 0.84, $P < 2.2 \times 10^{-16}$). All the anterior markers were negatively correlated with the pseudospace coordinate, whereas all posterior markers had a positive correlation (right).

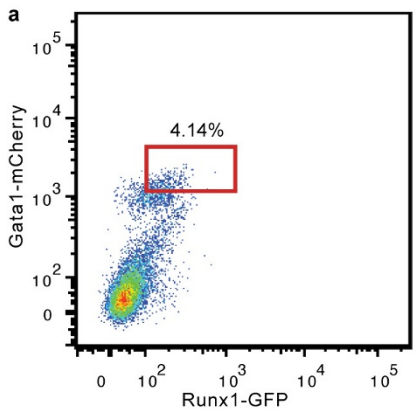


Extended Data Figure 7 | Expression of key genes along the anterior–posterior axis of the primitive streak in E7.0–7.75 embryos. Schematic representations of gene expression were generated from published *in situ* hybridization data (see citations) for key markers of clusters 4 (blue, mesoderm) and 7 (yellow, posterior mesoderm/blood progenitors). Expression of *T* (Brachyury)[97] and *Flk1* (Kdr, from in-house data) are shown to illustrate the extent of the primitive streak at E7.5. *Lefty2* and *Tbx6* (ref. 59) are expressed in the putative anterior portion of cluster 4 and in more anterior regions of the primitive streak in *in situ* analysis. *Tbx3* (ref. 98) and *Bmp4* (ref. 99) are expressed in the more posterior portion of cluster 4 and in the embryo are expressed in the more posterior region of the primitive streak around the amnion and into the extra-embryonic mesoderm. *Tek* and *Fli1* (from in-house data) are expressed in cluster 7 and in the embryo are found exclusively in the extra-embryonic portion. Also shown is the t-SNE for the cells from E7.0 onwards (primitive streak, neural plate and head fold stages) indicating expression of each gene (white, low; purple, high).



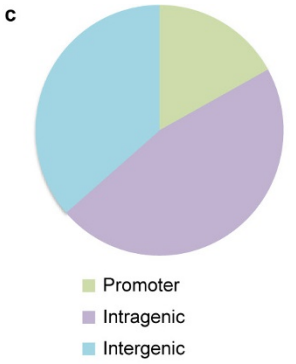
Extended Data Figure 8 | See next page for caption.

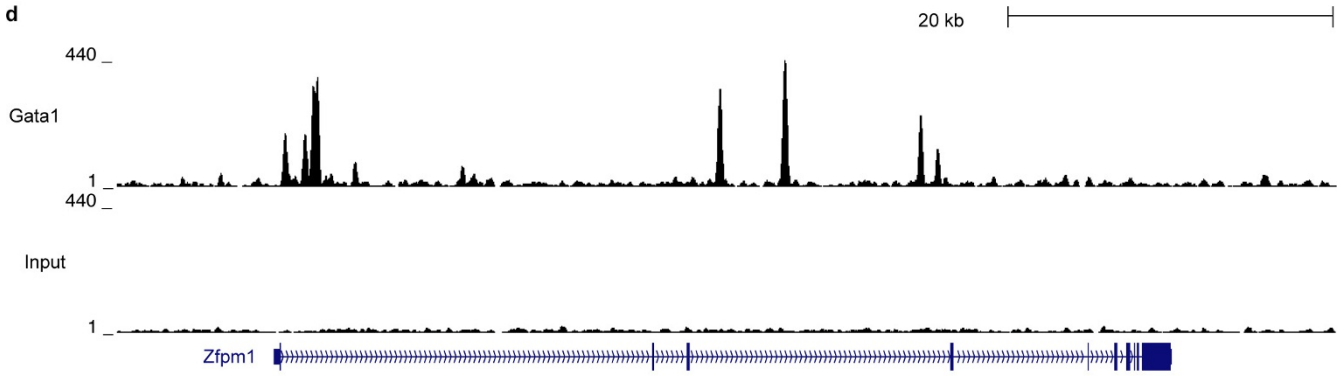
Extended Data Figure 8 | Pseudotime analysis of primitive erythroid development. a, Diffusion plot of the 271 cells in clusters 7 and 8. Different colours correspond to different plates and lanes of flow cells. **b**, Table showing the number of cells in each stage collected on the different plates (S, primitive streak; NP, neural plate; HF, head fold). **c**, Analogously to Extended Data Fig. 6, the angle α identifies a direction in the diffusion space (left panel). The percentage of variance explained by the batch effect associated to plates SLX-8344 and SLX-8345 is plotted as a function of α in the right panel. **d**, The pseudotime coordinate is robust to the use of different dimensionality reduction techniques, as shown in the left panel with t-SNE (Spearman correlation 0.92, $P < 2.2 \times 10^{-16}$) and in the right panel with independent component analysis (Spearman correlation 0.97, $P < 2.2 \times 10^{-16}$; same procedure described in Extended Data Fig. 6d). **e**, Plots showing the average expression of genes in clusters 1–3 of Fig. 4c along the pseudotime axis. Gene expression levels are normalized between 0 and 1. Dark red lines are the average expression levels of genes in each cluster as obtained from the fitting procedure, after normalization. Red shaded areas indicate standard deviation. **f**, Principal component analysis was performed on the expression pattern of genes known from previous studies to be upregulated or downregulated along the blood developmental trajectory[15,66,100–104]. The first principal component (explaining 44% of total variance) showed a very strong correlation with the pseudotime coordinate (left; Spearman correlation coefficient 0.91, $P < 2.2 \times 10^{-16}$). All upregulated (downregulated) genes positively (negatively) correlate with the pseudotime coordinate (right).

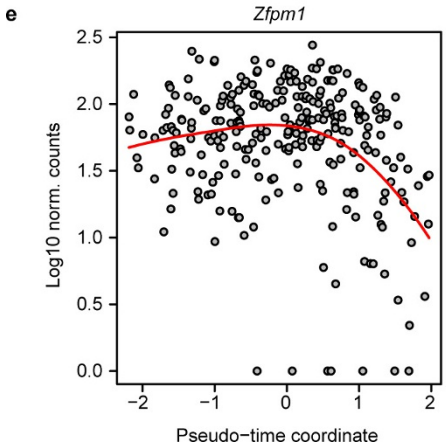


b

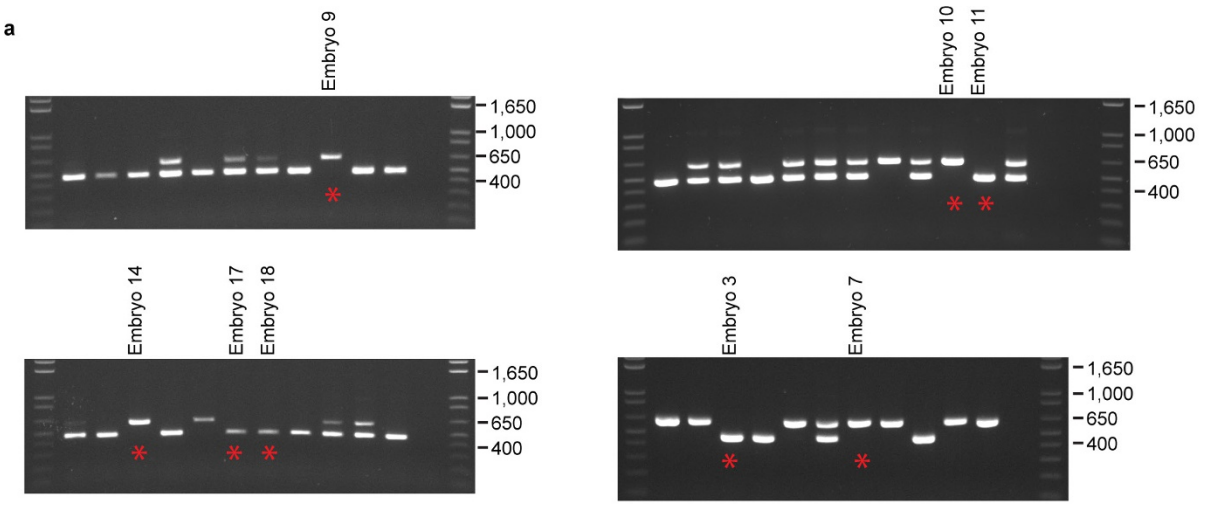
	Reads	Uniquely mappable reads	Peaks
Input	60,686,915	49,696,078	-
Gata1	81,003,777	52,522,929	4,135

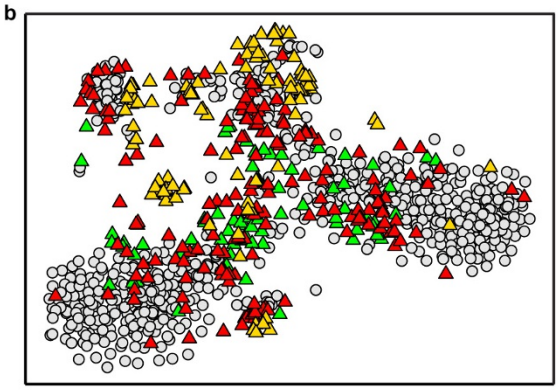


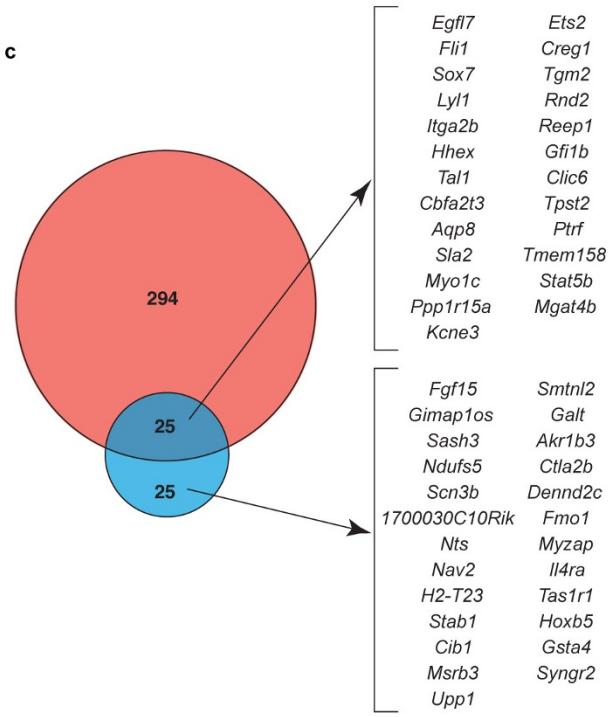




Extended Data Figure 9 | ChIP-seq for Gata1 in ESC-derived haematopoietic cells. a, Flow cytometry for Gata1–mCherry and Runx1-IRES–GFP knock-in reporter genes in embryoid body cells after 5 days of haematopoietic differentiation. Cells were sorted for the expression of both Runx1-IRES–GFP and Gata1–mCherry knock-in reporter genes to provide *in vitro* equivalents of the developing primitive erythrocytes assayed by RNA-seq. The gate used for sorting is shown in red. **b**, Numbers of reads and peaks identified for Gata1 and an input sample after mapping and peak calling; 4,135 Gata1 peaks were identified. **c**, Distribution of Gata1 peaks between promoter, intragenic and intergenic sequences. **d**, University of California, Santa Cruz Genome Browser tracks for Gata1 and input sample at the *Zfpm1* (Fog1) locus known to be a target of Gata1, indicating the quality of the ChIP-seq data. **e**, Expression of Gata1 target *Zfpm1* during the pseudotimecourse for erythroid development, as in Fig. 4.







d

	p-value	adjusted p-value
Tal1 HPC7 Progenitor Model (Wilson 2010)	0.962578	1
Tal1 Haemangioblast (Goode 2016)	0.006961	0.02088
Tal1 Haemogenic Endothelium (Goode 2016)	4.31E-05	0.00017
Tal1 Haematopoietic Progenitor (Goode 2016)	0.617238	1

Extended Data Figure 10 | Collection of embryos from *Tal1 LacZ/+* crosses. a, Genotyping PCR for embryos from *Tal1 LacZ/+* crosses. Lower band is the WT allele and upper band is the mutant allele carrying a neomycin knock in. Presence of both bands indicates heterozygosity. Embryos from which sequencing data were obtained are indicated with a red star and the number given corresponds to embryo identity in the metadata available online with the sequencing data. **b**, t-SNE as in Fig. 5d showing *Tal1* data (triangles; 377 cells) and original WT data (grey circles; 1,205 cells). *Tal1* data are coloured according to the embryo stage from which they were collected: green, neural plate; red, head fold; orange, four-somite pair. **c**, As in Fig. 5d, showing the complete list of genes. **d**, Gene set control analysis[105] was used to identify statistically significant overlaps between genes significantly downregulated in *Tal1*$^{-/-}$ compared with WT cells in the endothelial cluster (see Fig. 5) and Tal1 targets identified by ChIP-seq. Gene set control analysis identified an enrichment of our gene set with Tal1 ChIP-seq in ESC-derived haemangioblasts and haemogenic endothelium[106], but not in ESC-derived haematopoietic progenitors[106] or a haematopoietic progenitor cell line[52].