ARTICLE

https://doi.org/10.1038/s41586-019-0933-9

A single-cell molecular map of mouse gastrulation and early organogenesis

Blanca Pijuan-Sala[1,2,11], Jonathan A. Griffiths[3,11], Carolina Guibentif[1,2,11], Tom W. Hiscock[3,4], Wajid Jawaid[1,2], Fernando J. Calero-Nieto[1,2], Carla Mulas[2], Ximena Ibarra-Soria[3], Richard C. V. Tyser[5], Debbie Lee Lian Ho[2], Wolf Reik[6,7,8], Shankar Srinivas[5], Benjamin D. Simons[2,4,9], Jennifer Nichols[2], John C. Marioni[3,8,10]* & Berthold Göttgens[1,2]*

Across the animal kingdom, gastrulation represents a key developmental event during which embryonic pluripotent cells diversify into lineage-specific precursors that will generate the adult organism. Here we report the transcriptional profiles of 116,312 single cells from mouse embryos collected at nine sequential time points ranging from 6.5 to 8.5 days post-fertilization. We construct a molecular map of cellular differentiation from pluripotency towards all major embryonic lineages, and explore the complex events involved in the convergence of visceral and primitive streak-derived endoderm. Furthermore, we use single-cell profiling to show that *Tal1*$^{-/-}$ chimeric embryos display defects in early mesoderm diversification, and we thus demonstrate how combining temporal and transcriptional information can illuminate gene function. Together, this comprehensive delineation of mammalian cell differentiation trajectories in vivo represents a baseline for understanding the effects of gene mutations during development, as well as a roadmap for the optimization of in vitro differentiation protocols for regenerative medicine.

The 48 h of mouse embryonic development from embryonic day (E) 6.5 to E8.5 encompass the key phases of gastrulation and early organogenesis, when pluripotent epiblast cells diversify into ectodermal, mesodermal and endodermal progenitors of all major organs[1]. Despite the central importance of this period of mammalian development, we currently lack a comprehensive understanding of the underlying developmental trajectories and molecular processes involved, principally because previous research efforts have used in vitro systems[2], focused on small numbers of genes, or limited the number of developmental stages or cell types that were studied[3,4].

A single-cell map of early embryogenesis

To investigate the dynamic process of cellular diversification during gastrulation and early organogenesis, we complemented a previous E8.25 dataset[5] by generating single-cell RNA sequencing (scRNA-seq) profiles from 411 whole mouse embryos that were collected at six-hour intervals between E6.5 and E8.5 (Fig. 1a, b; Extended Data Figs. 1, 2a). Our dataset thus captures Theiler stages (TS) 9, TS10, TS11, and TS12, which are enriched in the pre-streak to early streak, mid-streak to late-streak, neural plate, and headfold to somitogenesis stages, respectively[6].

A total of 116,312 single-cell transcriptomes passed stringent quality control measures, with a median of 3,436 genes detected per cell (Methods; Extended Data Fig. 2b–d; Supplementary Table 1). We used clustering and cell annotation to identify 37 major cell populations (Methods; Fig. 1c; Extended Data Fig. 2e), and the presence of these cell types was associated with progression along the densely sampled time points (Extended Data Fig. 3a–d). The frequency of pluripotent epiblast cells declined over time, and mesodermal and definitive endodermal lineages appeared as early as E6.75. From E7.5 onwards, ectodermal lineages emerged alongside a diversification of cell types from each germ layer at the onset of organogenesis (Fig. 1d).

The transcriptional similarities that we found between clusters (Methods; Extended Data Fig. 3e, f) were in accordance with prior knowledge: epiblast was similar to neuroectoderm and primitive streak, and primitive streak was related to mesoderm and endoderm, consistent with the divergence of the three germ layers. Neural and mesodermal layers were connected during organogenesis (E8.25–E8.5) via a neuromesodermal progenitor population, which has been reported to give rise to both trunk mesoderm and neural tissues of the spinal cord[7,8] (Extended Data Fig. 3e). Our atlas can be explored via the interactive website: https://marionilab.cruk.cam.ac.uk/MouseGastrulation2018/.

Mapping endoderm development

Previous lineage-tracing studies[3,9] have shown that extra-embryonic and intra-embryonic endodermal cells intercalate to form a single tissue, highlighting the plasticity of embryonic cells (Extended Data Fig. 4a). Because we sampled extra-embryonic structures alongside the gastrulating embryo, our dataset provided an opportunity to investigate this convergence of primitive streak-derived definitive endoderm with visceral endoderm-derived cells at the molecular level.

To this end, we performed a focused analysis using only the visceral endoderm, anterior primitive streak, definitive endoderm, and gut cell types (Fig. 2a; 5,015 cells). The results were consistent with the gut endoderm arising from visceral as well as definitive endoderm (identified by expression of *Ttr* and *Mixl1*, respectively; Extended Data Fig. 4b, c). Inspection of the time points of cell collection supported the transcriptional convergence of these two lineages during development (Fig. 2b).

To define the transcriptional diversity within the maturing gut, we exclusively analysed cells collected at E8.25 and E8.5 and identified seven clusters that corresponded to different cell populations that line the gut tube (Fig. 2c). These spanned the pharyngeal endoderm

[1]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK. [2]Wellcome-Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. [3]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [4]The Wellcome/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, UK. [5]Department of Physiology Anatomy and Genetics, University of Oxford, Oxford, UK. [6]Epigenetics Programme, Babraham Institute, Cambridge, UK. [7]Centre for Trophoblast Research, University of Cambridge, Cambridge, UK. [8]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK. [9]Cavendish Laboratory, Department of Physics, University of Cambridge, Cambridge, UK. [10]EMBL-European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK. [11]These authors contributed equally: Blanca Pijuan-Sala, Jonathan A. Griffiths, Carolina Guibentif. *e-mail: john.marioni@cruk.cam.ac.uk; bg200@cam.ac.uk

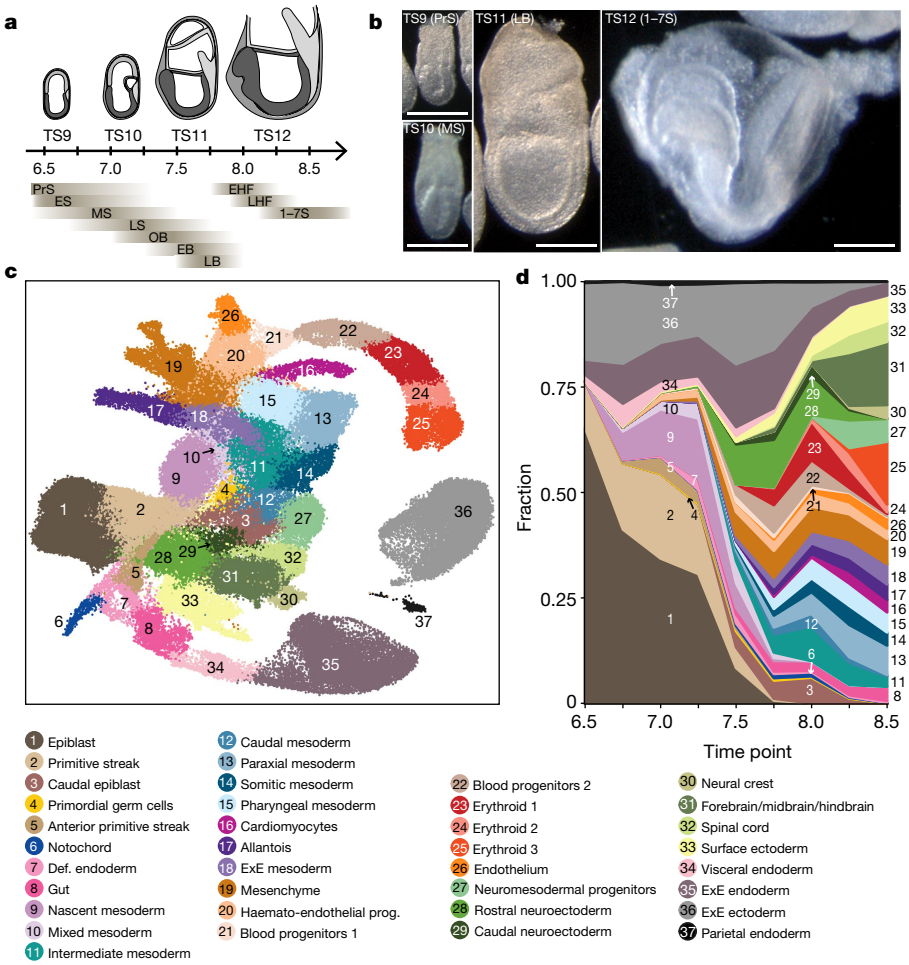


Fig. 1 | A single-cell resolution atlas of mouse gastrulation and early organogenesis. **a**, Overview of embryonic developmental time points sampled, alongside corresponding Theiler stages (TS9–TS12) and Downs and Davies stages. Adapted from a previous publication[38]. Numbers indicate days post-fertilization. PrS, pre-streak; ES, early streak; MS, mid-streak; LS, late streak; OB, neural plate no bud; EB, neural plate early bud; LB, neural plate late bud; EHF, early headfold; LHF, late headfold; 1–7S, 1–7 somites. **b**, Representative images of sampled embryos (see Supplementary Table 1 for sample collection details). Scale bars, 0.25 mm. **c**, Uniform manifold approximation and projection (UMAP) plot showing all the cells of the atlas (116,312 cells). Cells are coloured by their cell-type annotation and numbered according to the legend below. Def., definitive; ExE, extra-embryonic; prog., progenitor. **d**, Fraction of cell type per time point, displaying a progressive increase in cell-type complexity throughout our sampling.

(expressing *Nkx2-5*), foregut (expressing *Pyy*), midgut (expressing *Nepn*), and hindgut (expressing *Cdx2*) (Extended Data Fig. 4d–g). Notably, the foregut was split into two clusters, foregut 1 and foregut 2, which are likely to correspond to liver and lung precursors, respectively (see hepatic-associated genes *Hhex*, *Sfrp5*, and *Ttr*[10–12] and lung-associated genes *Ripply3* and *Irx1*[13,14] in Fig. 2d). Hindgut cells were also divided into two distinct clusters, hindgut 1 and hindgut 2, with significantly higher expression of the X chromosome genes *Trap1a* and *Rhox5* in hindgut 1 (Fig. 2d, e). Given the spatial complexity of the gut tube, we derived a pseudospatial ordering of these clusters using diffusion pseudotime (DPT[15]), which recapitulated their anterior–posterior distribution (Fig. 2f).

To assess how visceral endoderm cells might contribute to the maturing gut, we inferred cellular transitions along sequential collection time points using transport maps[16] (Methods). We then asked whether cells in each gut cluster at E8.25 and E8.5 were more likely to be derived from E7.0 visceral endoderm or definitive endoderm ancestors. To account for cells with a permanent extra-embryonic fate, we added extra-embryonic (ExE) endoderm to the analysis (Methods; Extended Data Fig. 4h). Both the ExE endoderm and the hindgut 1 clusters consisted of cells predicted to derive primarily from the visceral endoderm, whereas all remaining clusters were inferred to be predominantly of definitive endoderm origin (Fig. 2g; Extended Data Fig. 4i). Of note, the hindgut 1-specific genes *Trap1a* and *Rhox5* were also expressed in ExE endoderm and ExE ectoderm, in keeping with the extra-embryonic origin of the hindgut 1 cluster (Extended Data Fig. 4j). This suggests that, although hindgut 1 and hindgut 2 share a core hindgut signature, hindgut 1 cells also retain a transcriptional legacy from their extra-embryonic origin. We also extended previous work that examined lineage tracing at E8.75[9], by using a *Ttr::Cre* transgene coupled with a conditional yellow fluorescent protein (YFP) transgene in the *ROSA26* locus to show that Ttr–YFP-traced cells were enriched in the most posterior section of E8.5 embryos (Extended Data Fig. 4k).

Next, we inferred which cells belonged to the developmental trajectories from the visceral endoderm to hindgut 1, and from the definitive endoderm to hindgut 2 (Methods; Fig. 2h; Extended Data Fig. 5a, b); ordered the cells using DPT[15]; and clustered genes based on their expression dynamics along each trajectory (Methods; Supplementary Table 2; Extended Data Fig. 5c). We divided each trajectory into two domains, the first corresponding to gene expression before completion of endoderm intercalation at E7.5, and the second to gene expression after this point[17]. In the visceral endoderm–hindgut 1 trajectory, we observed the upregulation of visceral endoderm genes within the first domain, followed by an abrupt decline as cells proceeded towards the gut fate (Extended Data Fig. 5d). This suggests that we captured a subset of visceral endoderm cells that were undergoing visceral maturation before the onset of definitive endoderm intercalation.

Across both trajectories, a common set of genes was upregulated during intercalation (Extended Data Fig. 5c, e). This set included genes involved in epithelial remodelling such as *Pcna*, *Epcam*, and *Vim*, which is consistent with the epithelial arrangement expected to happen at this stage[9]. Genes commonly upregulated during the subsequent gut maturation and morphogenesis phase (Extended Data Fig. 5c, f) were enriched for transcription factors (over 20% of overlapping genes), and 66% of these were homeodomain proteins that showed sequential activation profiles, indicative of a temporal collinearity during hindgut specification[18]. Analysis of dynamic gene expression also revealed transcription factors that were specifically induced early in the visceral endoderm–hindgut 1 trajectory, including *Hes1*, *Pou5f1*, and *Sox4*. These represent promising candidates for further study (Extended Data Fig. 5g).

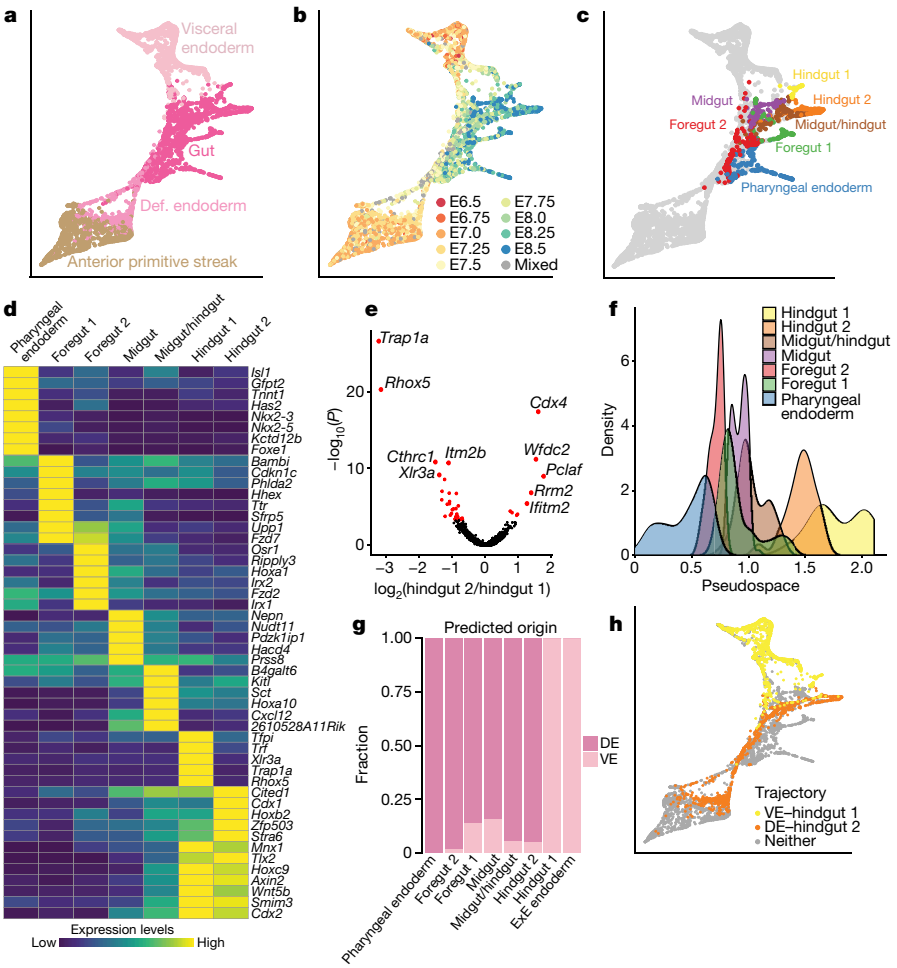


Fig. 2 | Molecular conversion and subsequent diversification during early endoderm development. a–c, Force-directed graph layout of the endoderm cell subset (5,015 cells) coloured by global cell-type annotation (**a**), embryo collection time point (**b**), or maturing gut cell type (**c**). Each point represents a cell, and cells close to each other have similar transcriptional profiles. **d**, Heat map illustrating the row-normalized mean expression of marker genes for each maturing gut cluster. **e**, Volcano plot showing the differentially expressed genes between hindgut 1 ($n = 53$ cells) and hindgut 2 ($n = 148$ cells). The x axis shows $\log_2$(gene expression in hindgut 2/gene expression in hindgut 1). Genes that are in red were differentially expressed to a significant level (Benjamini–Hochberg-adjusted $P < 0.1$; Methods). The five most significantly differentially expressed genes in each direction are labelled. **f**, Pseudospatial ordering of cells along the gut tube. Pseudospace coordinates (x axis) correspond to diffusion pseudotime (DPT) values. **g**, Fraction of cells from each maturing gut cluster that are predicted to derive from visceral endoderm (VE) or definitive endoderm (DE). **h**, Force-directed graph coloured by putative trajectories for formation of the hindgut clusters.

Origins of haemato–endothelial lineages

Red blood cells are formed in two consecutive waves in the yolk sac; the first arising at around E7.5 and the second from E8.25[19]. The first wave (primitive) generates nucleated erythrocytes, which disappear shortly after birth. The second wave (yolk-sac-definitive) starts with the emergence of erythro-myeloid progenitors (EMPs) from yolk-sac haemogenic endothelium. These later migrate to the fetal liver and generate definitive erythrocytes[19] (Extended Data Fig. 6a).

Although some key phenotypic and molecular distinctions between primitive and yolk-sac-definitive haematopoiesis are known, the respective in vivo progenitor cells are poorly understood owing to limiting cell numbers and lack of markers. To characterize these processes in more depth, we computationally isolated and re-clustered cells assigned to the erythroid, haemato-endothelial, blood progenitor, endothelial, and mixed mesoderm groups (15,875 cells; Fig. 3a, b; Extended Data Fig. 6b).

This analysis highlighted a putative trajectory towards the primitive erythroid lineage, passing through the sub-clusters haemato-endothelial progenitors 1 and 2 (Haem 1 and Haem 2), blood progenitors 1 and 2 (BP1 and BP2), and erythroid 1–4 (Ery 1–Ery 4) (Fig. 3a, b). This trajectory did not include the endothelial region (sub-clusters EC1–EC8), which was enriched for cells collected at E8.25–E8.5, displayed a complex structure, and expressed high levels of *Kdr*, which encodes the protein FLK1 (hereafter, we term this the *Kdr*^{hi} region (Fig. 3c)). Notably, some of these endothelial cells expressed haematopoietic markers, such as *Spi1* (also known as *PU.1*) and *Itga2b* (Fig. 3d), potentially highlighting the emergence of blood from endothelium during the second wave[20]. The incorporation of temporal information (Fig. 3b) suggested that, unlike the second haematopoietic wave, the first wave does not transit through a molecular state with classical characteristics of mature endothelial cells, such as high expression of *Cdh5* and *Pecam1* (Extended Data Fig. 6c).

Endothelial cells are generated independently in the yolk sac, allantois and embryo proper[21,22], and the allantoic endothelium is hypothesized to display a specific transcriptional signature[5]. To test whether the heterogeneity in the *Kdr*^{hi} region was associated with different anatomical locations, we dissected out the yolk sac, allantois, and embryo proper from a new batch of E8.25 embryos, purified the endothelial cells by flow sorting the FLK1$^+$ population, and performed scRNA-seq on 288 cells using Smart-seq2[23] (Extended Data Fig. 6d–f). Assigning the cells from the *Kdr*^{hi} region (sub-clusters EC1–EC8) to their most likely embryonic location suggested that diverse anatomical origin could partially explain the transcriptional heterogeneity observed in the endothelium (Fig. 3e).

Previous in vitro colony-forming assays of early embryonic cells suggested that, in addition to erythrocytes, the primitive wave also gives rise to macrophage and megakaryocytic progenitors[24–26]. However, the molecular nature of these progenitors remains obscure. In our atlas, we identified two rare cell groups (present at a frequency of around 0.1%) that we annotated as megakaryocytes and myeloid cells (Fig 3f; Extended Data Fig. 6g). This provided us with an opportunity to characterize their molecular profiles based on primary in vivo cells.

Previous reports suggest that early myeloid progenitors can give rise to brain microglia[27]. Consistent with this, cells in our myeloid cell population expressed *Ptprc* (encoding CD45), *Kit*, *Csf1r*, and *Fcgr3* (encoding CD16),which were previously reported to be markers of the E8.5 EMP-like population that gives rise to microglial macrophages[20,28] (Fig. 3g). However, we did not detect the more mature microglial-related genes such as *Cx3cr1*, *Adgre1* (encoding F4/80), and only saw low levels of *Tmem119*[29,30]. To investigate the location and frequency of these cells in the embryo, we dissected different regions of E8.5 embryos

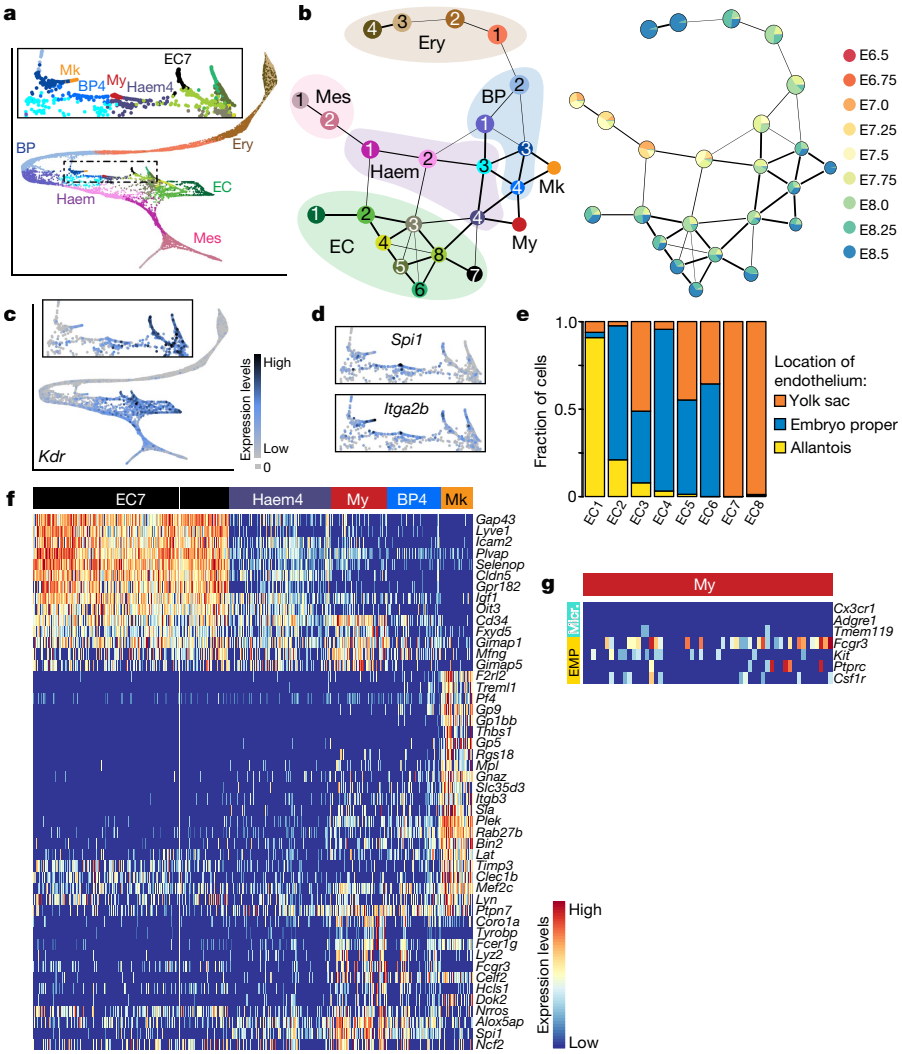


Fig. 3 | Temporal analysis of blood emergence reveals early myeloid cells. a, Force-directed graph layout of cells associated with the blood lineage, coloured by sub-cluster (15,875 cells). The inset box shows a zoomed-in section that focuses on myeloid, megakaryocytic, and haemogenic endothelial cells. BP, blood progenitor; EC, endothelial cell; Ery, erythrocyte; Haem, haemato-endothelial progenitor; Mes, mesodermal cell; Mk, megakaryocyte; My, myeloid cell. **b**, Graph abstraction summarizing the relationships between the sub-clusters as in **a**, coloured by sub-cluster (left) and collection time point (right). Two samples of mixed-time point embryos were excluded. **c**, Expression levels of *Kdr*, overlaid on the force-directed layout from **a**. **d**, Expression levels of *Spi1* and *Itga2b*, overlaid on the inset of the force-directed layout from **a**. **e**, Fraction of endothelial cells that mapped to yolk sac, allantois, and embryo proper. **f**, Heat map illustrating row-normalized expression of genes that were significantly upregulated in cells of the EC7 ($n = 197$), Haem4 ($n = 102$), My ($n = 56$), BP4 ($n = 54$), and Mk ($n = 32$) sub-clusters when performing pairwise differential expression analyses between a specific sub-cluster and the rest of the cells in **a**. Significance was considered if $\log_2$(mean expression of specific cluster/mean expression of the rest of cells) > 2.5 and Benjamini–Hochberg-adjusted $P < 0.05$. **g**, Heat map illustrating the log-count expression ($\log_2$(normalized count + 1), ranging from 0 (blue) to 3.5 (red)) of previously described microglial (Micr.) and erythro-myeloid progenitor (EMP) markers.

(Extended Data Fig. 6h) and performed flow cytometry analysis using the markers CD16/32, and CSF1R. Rare CD16/32$^+$CSF1R$^+$ cells were found in all dissected regions (Extended Data Fig. 6i), indicating that by E8.5 this population has already started to migrate out of the yolk sac.

A platform to dissect genetic mutations

Previous work has emphasised the critical role of the basic helix–loop–helix (bHLH) transcription factor TAL1 (also known as SCL) in haematopoiesis; in these experiments, *Tal1*$^{-/-}$ mouse embryos died of severe anaemia at around E9.5[31]. Dissecting the temporal and mechanistic roles of such major regulatory genes in vivo is challenging using knockout mice—breeding mice and genotyping embryos is time-consuming, and furthermore, the direct effects of a mutation are often masked by gross developmental malformations or embryo lethality. To circumvent these difficulties, we generated chimeric mouse embryos in which *Tal1*$^{-/-}$ tdTomato$^+$ mouse embryonic stem (ES) cells were injected into wild-type blastocysts. In the resulting chimaeras, wild-type cells still produce blood cells, and this allows the specific effects of TAL1 depletion to be studied in an otherwise healthy embryo[32].

To determine whether *Tal1* mutant cells were associated with abnormalities in specific lineages, we sorted tdTomato$^-$ (wild type) and tdTomato$^+$ (*Tal1*$^{-/-}$) cells from chimeric embryos at E8.5, and then performed scRNA-seq (Fig. 4a; Extended Data Fig. 7a, b). Each cell was annotated by computationally mapping its transcriptome onto our wild-type atlas (Methods; Fig. 4b; Extended Data Fig. 7c–e). Consistent with the pivotal role of *Tal1* in haematopoiesis, tdTomato$^+$ cells did not contribute to blood lineages (Fig. 4b; Extended Data Fig. 7e–g). Notably, we confirmed that wild-type control tdTomato$^+$ *Tal1*$^{+/+}$ ES cells, when injected into wild-type embryos, make a similar contribution to haematopoiesis as the tdTomato$^-$ host cells (Extended Data Fig. 7h, i).

Comparisons between wild-type and *Tal1*$^{-/-}$ chimeric cells mapped to the landscape defined in Fig. 3a illustrated that TAL1 depletion

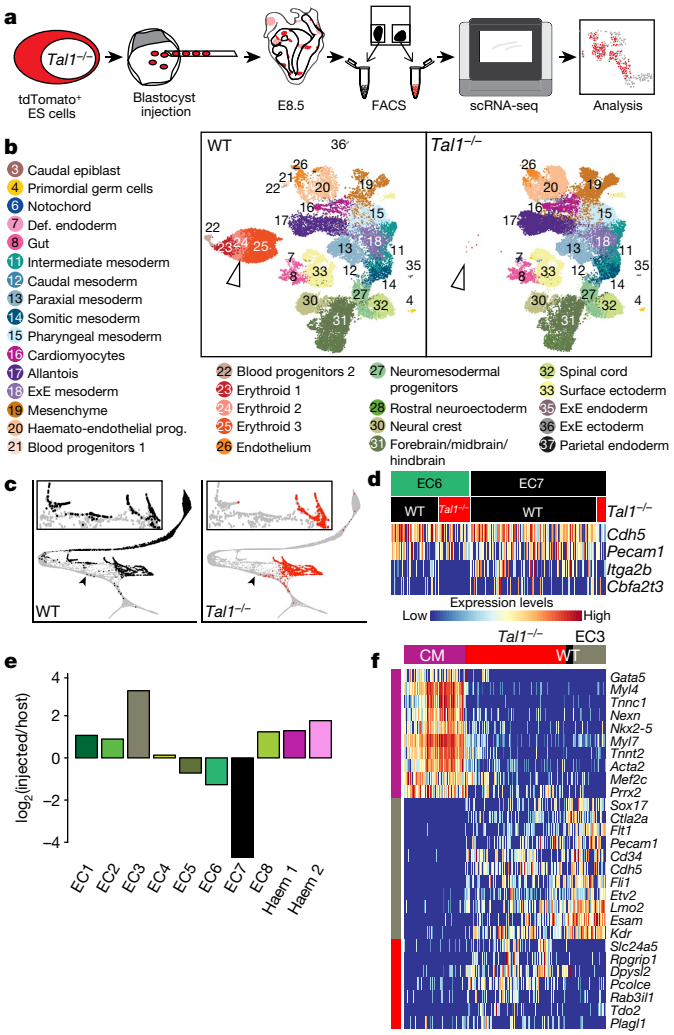


Fig. 4 | Mapping *Tal1*$^{-/-}$ chimaeras to the atlas identifies molecular states associated with defects in haemato-endothelial development. **a**, Experimental design for *Tal1*$^{-/-}$ chimaera generation and sequencing. FACS, fluorescence-activated cell sorting. **b**, UMAP plots of chimaera cells ($n = 25{,}078$ wild type (WT); $n = 26{,}326$ *Tal1*$^{-/-}$). Points are coloured and numbered according to their computationally assigned cell type, as in Fig. 1. Numbers and legend are only specified for those cell types that are clearly visible in the plot. White arrowheads indicate blood cells (which are depleted in mutant cells). **c**, Mapping of blood-related cells from the chimaera onto the blood-related cells from the atlas. Left: wild type ($n = 9{,}336$); right: *Tal1*$^{-/-}$ ($n = 2{,}911$). Black arrowheads denote the position at which blood development appears to be blocked in *Tal1*$^{-/-}$ cells. **d**, Heat map illustrating the row-normalized expression of blood (*Cbfa2t3* and *Itga2b*) and endothelial (*Cdh5* and *Pecam1*) genes in EC6 wild type ($n = 43$), EC6 *Tal1*$^{-/-}$ ($n = 28$), EC7 wild type ($n = 117$), and EC7 *Tal1*$^{-/-}$ ($n = 7$) cells. **e**, Change in abundance of *Tal1*$^{-/-}$ cells with respect to wild-type chimaera cells in each of the sub-clusters ($\log_2$(mapped *Tal1*$^{-/-}$ cell fraction/mapped wild-type cell fraction)). **f**, Heat map illustrating the row-normalized expression of genes that were upregulated in EC3-mapped *Tal1*$^{-/-}$ cells. From left to right, columns represent a sample of atlas cardiomyocytes (CM) ($n = 200$), *Tal1*$^{-/-}$ EC3 ($n = 328$), wild type EC3 ($n = 23$), and atlas EC3 ($n = 107$) cells. Illustrative genes were manually selected from the full heat map, which is shown in Extended Data Fig. 8a.

disrupts the emergence of primitive erythroid cells as well as our newly characterized megakaryocyte and myeloid cells (Fig. 4c). Although a subset of *Tal1*$^{-/-}$ cells were mapped to the haemogenic endothelial groups EC6 and EC7, they were defective in the expression of genes associated with blood development, such as *Itga2b* or the known TAL1 target gene *Cbfa2t3* (also known as *Eto2*), in contrast with the host wild-type cells. These findings suggest that the second haematopoietic wave is also disrupted after TAL1 depletion (Fig. 4d).

To characterize this developmental block in the second haematopoietic wave further, we quantified the relative contributions of *Tal1*$^{-/-}$ and wild-type chimeric cells to each endothelial (EC1–EC8) and each haemato-endothelial (Haem 1 and Haem 2) cluster described in Fig. 3 (Fig. 4e; Supplementary Table 3). Of note, E8.5 *Tal1*$^{-/-}$ cells were more abundant than wild-type cells in EC3, one of the earliest-appearing endothelial sub-clusters (Fig. 3b). While mutant cells might simply accumulate in this state, *Tal1*$^{-/-}$ cells mapped to EC3 may alternatively acquire a transcriptional state that is similar but not identical to EC3. To clarify this, we performed differential expression analyses that compared EC3-mapped *Tal1*$^{-/-}$ cells firstly to their most similar cells in the reference atlas, and secondly to the wild-type host chimeric cells that were mapped to EC3. We observed a small number of genes that were specifically upregulated in EC3-mapped *Tal1*$^{-/-}$ cells, including *Pcolce*, *Tdo2*, and *Plagl1* (Fig. 4f; Extended Data Fig. 8a). When we examined the expression of these genes in our atlas, we observed high expression in the mesenchyme and other mesoderm clusters, such as the allantois, paraxial, pharyngeal, and intermediate mesoderm (Extended Data Fig. 8b). Furthermore, we noted that a subset of the EC3-mapped *Tal1*$^{-/-}$ cells also upregulated expression of cardiac-related genes such as *Nkx2-5*, *Mef2c*, and *Tnnt2* (Fig. 4f), consistent with a previous report that *Tal1*$^{-/-}$ yolk-sac cells can adopt a cardiomyocyte-like phenotype[33]. However, these cells did not display a full cardiomyocyte transcriptional program and continued to express endothelial genes such as *Esam* and *Sox17*, albeit with some downregulation compared with their wild-type EC3 atlas counterparts. These results suggest that *Tal1* disruption blocks cells at a transcriptional state similar to that of the EC3 sub-cluster during the second wave of blood development. Moreover, when unable to proceed towards a haemogenic phenotype, EC3-mapped *Tal1*$^{-/-}$ cells begin to activate other mesodermal programs. This is in accordance with prior evidence showing that haematopoietic precursors isolated from E7.5 mouse embryos are endowed with mesodermal plasticity when cultured ex vivo[34].

Discussion

Our comprehensive atlas of mouse gastrulation and early organogenesis offers a powerful resource for investigating the molecular underpinnings of cell-fate decisions during this key period of mammalian development. We exploited this resource by investigating two specific developmental phenomena: the transdifferentiation process of visceral endoderm cells that contribute to the composition of the embryonic gut; and the emergence of rare blood cells in the early embryo. Moreover, we used our atlas as a reference for the analysis of *Tal1*$^{-/-}$ mutant chimeric embryos, and highlighted where TAL1 is critical for progression into the blood lineage. We also identified a transcriptional state unique to *Tal1*$^{-/-}$ cells, within which genes from multiple different mesodermal tissues are expressed alongside endothelial genes.

More broadly, our chimaera analysis illustrates the utility and efficiency of such a model for studying the molecular and cellular consequences of a wide range of developmental mutants, including those that are embryonically lethal and relevant for human developmental disorders. Our work in the mouse, a widely used and experimentally relevant mammalian system, complements recent single-cell expression profiling surveys in early zebrafish and *Xenopus* embryos[35–37]. Collectively, these studies demonstrate that densely sampled large-scale single-cell profiling has the potential to advance our understanding of embryonic development in vertebrates.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-0933-9.

Received: 31 May 2018; Accepted: 20 December 2018;
Published online 20 February 2019.

1. Tam, P. P. L. & Behringer, R. R. Mouse gastrulation: the formation of a mammalian body plan. *Mech. Dev.* **68**, 3–25 (1997).
2. Loh, K. M. et al. Mapping the pairwise choices leading from pluripotency to human bone, heart, and other mesoderm cell types. *Cell* **166**, 451–467 (2016).
3. Viotti, M., Nowotschin, S. & Hadjantonakis, A.-K. SOX17 links gut endoderm morphogenesis and germ layer segregation. *Nat. Cell Biol.* **16**, 1146–1156 (2014).
4. Lescroart, F. et al. Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. *Science* **359**, 1177–1181 (2018).
5. Ibarra-Soria, X. et al. Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nat. Cell Biol.* **20**, 127–134 (2018).
6. Downs, K. M. & Davies, T. Staging of gastrulating mouse embryos by morphological landmarks in the dissecting microscope. *Development* **118**, 1255–1266 (1993).
7. Koch, F. et al. Antagonistic activities of *Sox2* and *Brachyury* control the fate choice of neuro-mesodermal progenitors. *Dev. Cell* **42**, 514–526.e7 (2017).
8. Tzouanacou, E., Wegener, A., Wymeersch, F. J., Wilson, V. & Nicolas, J.-F. Redefining the progression of lineage segregations during mammalian embryogenesis by clonal analysis. *Dev. Cell* **17**, 365–376 (2009).
9. Kwon, G. S., Viotti, M. & Hadjantonakis, A.-K. The endoderm of the mouse embryo arises by dynamic widespread intercalation of embryonic and extraembryonic lineages. *Dev. Cell* **15**, 509–520 (2008).
10. Finley, K. R., Tennessen, J. & Shawlot, W. The mouse *Secreted frizzled-related protein 5* gene is expressed in the anterior visceral endoderm and foregut endoderm during early post-implantation development. *Gene Expr. Patterns* **3**, 681–684 (2003).
11. Makover, A., Soprano, D. R., Wyatt, M. L. & Goodman, D. S. An in situ-hybridization study of the localization of retinol-binding protein and transthyretin messenger RNAs during fetal development in the rat. *Differentiation* **40**, 17–25 (1989).
12. Martinez Barbera, J. P. et al. The homeobox gene *Hex* is required in definitive endodermal tissues for normal forebrain, liver and thyroid formation. *Development* **127**, 2433–2445 (2000).
13. Bosse, A. et al. Identification of the vertebrate Iroquois homeobox gene family with overlapping expression during early development of the nervous system. *Mech. Dev.* **69**, 169–181 (1997).
14. Osipovich, A. B. et al. *Insm1* promotes endocrine cell differentiation by modulating the expression of a network of genes that includes *Neurog3* and *Ripply3*. *Development* **141**, 2939–2949 (2014).
15. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
16. Schiebinger, G. et al. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. *Preprint at* https://www.bioRxiv.org/content/early/2017/09/27/191056 (2017).
17. Viotti, M., Foley, A. C. & Hadjantonakis, A. K. Gutsy moves in mice: cellular and molecular dynamics of endoderm morphogenesis. *Phil. Trans. R. Soc. Lond. B* **369**, 20130547 (2014).
18. Deschamps, J. & Duboule, D. Embryonic timing, axial stem cells, chromatin dynamics, and the Hox clock. *Genes Dev.* **31**, 1406–1416 (2017).
19. Palis, J. Hematopoietic stem cell-independent hematopoiesis: emergence of erythroid, megakaryocyte, and myeloid potential in the mammalian embryo. *FEBS Lett.* **590**, 3965–3974 (2016).
20. McGrath, K. E. et al. Distinct sources of hematopoietic progenitors emerge before HSCs and provide functional blood cells in the mammalian embryo. *Cell Reports* **11**, 1892–1904 (2015).
21. Downs, K. M., Gifford, S., Blahnik, M. & Gardner, R. L. Vascularization in the murine allantois occurs by vasculogenesis without accompanying erythropoiesis. *Development* **125**, 4507–4520 (1998).
22. Patan, S. in *Angiogenesis in Brain Tumors* (eds Kirsch, M. & Black, P. M.) 3–32 (Springer, Boston, MA, 2004).
23. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
24. Palis, J., Robertson, S., Kennedy, M., Wall, C. & Keller, G. Development of erythroid and myeloid progenitors in the yolk sac and embryo proper of the mouse. *Development* **126**, 5073–5084 (1999).
25. Tober, J. et al. The megakaryocyte lineage originates from hemangioblast precursors and is an integral component both of primitive and of definitive hematopoiesis. *Blood* **109**, 1433–1441 (2007).
26. Xu, M.-j. et al. Evidence for the presence of murine primitive megakaryocytopoiesis in the early yolk sac. *Blood* **97**, 2016–2022 (2001).
27. Hoeffel, G. et al. C-Myb^{+} erythro-myeloid progenitor-derived fetal monocytes give rise to adult tissue-resident macrophages. *Immunity* **42**, 665–678 (2015).
28. Gomez Perdiguero, E. et al. The origin of tissue-resident macrophages: when an erythro-myeloid progenitor is an erythro-myeloid progenitor. *Immunity* **43**, 1023–1024 (2015).
29. Bennett, M. L. et al. New tools for studying microglia in the mouse and human CNS. *Proc. Natl Acad. Sci. USA* **113**, E1738–E1746 (2016).
30. Ginhoux, F. et al. Fate mapping analysis reveals that adult microglia derive from primitive macrophages. *Science* **330**, 841–845 (2010).
31. 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).
32. Robb, L. et al. The *scl* gene product is required for the generation of all hematopoietic lineages in the adult mouse. *EMBO J.* **15**, 4123–4129 (1996).
33. Van Handel, B. et al. Scl represses cardiomyogenesis in prospective hemogenic endothelium and endocardium. *Cell* **150**, 590–605 (2012).
34. Huber, T. L., Kouskoff, V., Fehling, H. J., Palis, J. & Keller, G. Haemangioblast commitment is initiated in the primitive streak of the mouse embryo. *Nature* **432**, 625–630 (2004).
35. Briggs, J. A. et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018).
36. Farrell, J. A. et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131 (2018).
37. Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
38. Pijuan-Sala, B., Guibentif, C. & Göttgens, B. Single-cell transcriptional profiling: a window into embryonic cell-type specification. *Nat. Rev. Mol. Cell Biol.* **19**, 399–412 (2018).

Acknowledgements We thank W. Mansfield for blastocyst injections; A. T. L. Lun and F. Hamey for discussions concerning the analysis; T. L. Hamilton for technical support in embryo collection; S. Kinston and K. Jones for technical assistance; the Flow Cytometry Core Facility at CIMR for cell sorting; the CRUK-CI genomics core for the chimaera scRNA-seq 10x libraries and for letting us use the 10x Chromium after hours; the Wellcome Sanger Institute DNA Pipelines Operations for sequencing; and K. Hadjantonakis for sharing the *Ttr::cre* mouse line. Research in the authors' laboratories is supported by Wellcome, the MRC, CRUK, Bloodwise, and NIH-NIDDK; by core support grants from Wellcome to the Cambridge Institute for Medical Research and Wellcome-MRC Cambridge Stem Cell Institute; and by core funding from CRUK and the European Molecular Biology Laboratory. B.P.-S. and D.L.L.H. are funded by the Wellcome 4-Year PhD Programme in Stem Cell Biology and Medicine and the University of Cambridge; D.L.L.H. is also funded by the Cambridge Commonwealth European and International Trust. J.A.G. is funded by the Wellcome Mathematical Genomics and Medicine Programme at the University of Cambridge (109081/Z/15/A). C.G. is funded by the Swedish Research Council (2017-06278, administered by Sahlgrenska Cancer Center, University of Gothenburg). This work was funded as part of a Wellcome Strategic Award (105031/Z/14/Z) awarded to W.R., B.G., J.C.M., J.N., L. Vallier, S.S., B.D.S., S. Teichmann, and T. Voet; by a Wellcome grant (108438/Z/15) awarded to J.C.M. and S.S., and by a BBSRC grant (BBS/E/B/000C0421) awarded to W.R.

Reviewer information *Nature* thanks Peter Sims, Patrick Tam and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.P.-S., W.J., F.J.C.-N., C.M. and J.N. generated the atlas dataset. C.G. designed and executed the chimaera dataset generation and associated experiments. D.L.L.H. assisted in the generation of the *Tal1*$^{-/-}$ ES cell line. J.A.G. performed pre-processing, low-level analyses, batch correction, clustering, and global visualization of the atlas and chimaera datasets, and designed the associated website. B.P.-S. curated the clustering and evaluated the connectivity between cell types. B.P.-S. and C.G. annotated atlas cell types. J.A.G. and C.G. analysed atlas endoderm. B.P.-S. assisted in the endoderm analyses by generating force-directed layouts and inferring trajectories using graph abstraction as an alternative approach. R.C.V.T. performed *Ttr::cre* embryo imaging experiments. B.P.-S. analysed atlas haemato-endothelium and performed associated experiments and analyses. J.A.G. mapped chimaera cells to the atlas. B.P.-S. and C.G. analysed the effects of *Tal1*$^{-/-}$. T.W.H. contributed to the mapping and analysis of chimaeras. X.I.-S. provided advice on bioinformatics analysis. W.R., S.S., B.D.S., J.N., J.C.M., and B.G. supervized the study. B.P.-S., J.A.G., C.G., T.W.H., J.C.M., and B.G. wrote the manuscript. All authors read and approved the final manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-019-0933-9.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-0933-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.C.M. or B.G.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. The investigators were not blinded to allocation during experiments and outcome assessment.

Embryo collection and sequencing. All procedures were performed in strict accordance with the UK Home Office regulations for animal research. Chimeric mouse embryos were generated under the project licence number PPL 70/8406.

Reference atlas. Pregnant C57BL/6 female mice (mated at 7 weeks of age) were purchased from Charles River and delivered one day before or on the day of embryo harvest. Mouse embryos were dissected at time points E6.5, E6.75, E7.0, E7.25, E7.5, E7.75, E8.0, E8.25, and E8.5. As previously reported[6], development can proceed at different speeds between embryos, even within the same litter (Fig. 1a; Extended Data Fig. 1). Consequently, we adopted careful staging by morphology (Downs and Davies staging[6]) to exclude clear outliers. Following euthanasia of the females using cervical dislocation, the uteri were collected into PBS with 2% heat-inactivated FCS and the embryos were immediately dissected and processed for scRNA-seq. Two samples contained pooled embryos staged across several time points. Cells from these samples are denoted as 'mixed' in the figures, and 'mixed_gastrulation' in Supplementary Table 4. Embryos from the same stage were pooled to make individual 10x samples, and single-cell suspensions were prepared by incubating the embryos with TrypLE Express dissociation reagent (Life Technologies) at 37 °C for 7 min and quenching with heat-inactivated serum. The resulting single-cell suspension was washed and resuspended in PBS with 0.4% BSA, and filtered through a Flowmi Tip Strainer with 40-μm porosity (ThermoFisher Scientific, 136800040). Cell counts were then assessed with a haemocytometer. scRNA-seq libraries were subsequently generated using the 10x Genomics Chromium system (v.1 Chemistry) and samples were sequenced according to the manufacturer's instructions on an Illumina HiSeq 2500 platform. Detailed information on embryo collection is provided in Supplementary Table 1, and the metadata for each sequenced cell are provided in Supplementary Table 4. Sample sizes were chosen to maximize the number of recovered cells from each experiment and to obtain total cell numbers similar to the estimated cell numbers in mouse embryos at their respective stages. The sample sizes were also dependent on the number of viable embryos from each litter. When required, cells were partitioned into multiple samples to prevent overloading of a single 10x lane.

Yolk sac, allantois and embryo proper endothelial-cell dissection experiment. Mice were bred and maintained at the University of Cambridge, in individually ventilated cages with sterile bedding; sterile food and water were provided ad libitum. All animals were kept in pathogen-free conditions. Timed matings were set up between C57BL/6 mice, purchased from Charles River. Upon dissection, only embryos staged as TS12 were further processed. Allantois, yolk sac, and embryo proper were dissected and placed into separate tubes. Single-cell suspensions were prepared by incubating the embryos with TrypLE Express dissociation reagent (Life Technologies) at 37 °C for 7 min and quenching with heat-inactivated serum. Single cells were subsequently stained with FLK1-PE antibody (1:100; Biolegend, 12-5821-83, clone Avas12a1, E01819-1631) and 4′,6-diamidino-2-phenylindole (DAPI) as viability stain (1 μg/ml; Sigma). Live FLK1$^+$ cells were isolated by fluorescence-activated cell sorting (FACS) using a BD Influx sorter into individual wells of a 96-well plate containing lysis buffer (0.2% (v/v) Triton X-100 and 2 U/μl SUPERase-In (Invitrogen, AM2696)) and stored at −80 °C (1 plate per tissue was prepared). Plates were processed following the Smart-seq2 protocol as previously described[23] and libraries were generated using the Illumina Nextera XT DNA preparation kit. Libraries were pooled and sequenced on an Illumina HiSeq 4000. Sample sizes were chosen based on the number of viable endothelial cells recovered from the experiment. We also aimed to have an equal (or very similar) number of endothelial cells from each of the dissected regions, and a number that was large enough (that is, at least 70 per sample) to infer correlations with the atlas dataset.

Flow cytometry analysis of myeloid progenitors. Mice and embryos were obtained as described above. Yolk sac, allantois, amnion, head, heart, and trunk were dissected and placed into separate tubes. Single-cell suspensions were prepared as above and single cells were subsequently stained with CD16/32–BV711 (1:200; Biolegend, 101337, clone 93, B251800) for 20 min at 4 °C, washed with 2 ml PBS + 2% FCS, blocked with Fc block CD16/32 (1:100; eBioscience, 14-0161-85, clone 93, E03558-1640), and stained with CSF1R–BV605 (1:800; Biolegend; 135517, clone AFS98, B196541) for 30 min at 4 °C. Cells were then washed and 7-aminoactinomycin D (7-AAD) (1:200; BD Pharmigen; 51-68981E, 7061885) was added as a viability stain. Cells were analysed using a BD Fortessa cytometer. Gates were established using 'fluorescence minus one' controls. Two biological replicates were performed: one pool of 12 and one pool of 13 embryos.

Ttr–YFP embryo staining. *Ttr::cre* stud male mice[9] were crossed with R26R–YFP females[39]. Dissected E8.5 embryos were fixed for 1 h at room temperature with 4% paraformaldehyde (PFA) in PBS. The embryos were then washed 3 times in PBS with 0.1% Triton X-100 (PBT–0.1%) for 15 min, permeabilized in PBT–0.25% for 40 min, and washed again three times in PBT–0.1%. The embryos were transferred to blocking solution (5% donkey serum (Sigma, D9663), 1% BSA (Sigma, A7906) in PBT–0.1%) overnight at 4 °C. Primary antibody (chicken anti-GFP; 1:100; Abcam, ab13970, GR3190550-2) was then added to the blocking solution, and the samples were incubated in the solution overnight at 4 °C. The embryos were washed 3 times in PBT–0.1% and incubated overnight at 4 °C in PBT–0.1% with the secondary antibody (goat anti-chicken 488; Sigma; 1:100; A11039; 1899514) and phalloidin 555 (1:100; Sigma; #19083), then subsequently washed three times in PBT–0.1% for 15 min and mounted in Vectashield mounting media with DAPI for at least 24 h at 4 °C. Images were captured using a Zeiss 880 confocal microscope.

Chimaera generation and sequencing. TdTomato-expressing mouse ES cells were derived as previously described[40] from E3.5 blastocysts obtained by crossing a ROSA26–tdTomato male (Jax Labs; 007905) with a wild-type C57BL/6 female. The cells were negative for mycoplasma contamination. The cells were expanded under 2i + LIF conditions[41] and transiently transfected with a Cre-IRES–GFP plasmid[42] using Lipofectamine 3000 Transfection Reagent (ThermoFisher Scientific, L3000008) according to the manufacturer's instructions. Single GFP$^+$ cells were sorted 48 h after transfection into 96-well plates. Individual clones were allowed to grow and were manually picked for expansion. A tdTomato-positive, male, karyotypically normal line, competent for chimaera generation as assessed using morula aggregation assay, was selected for targeting *Tal1*. Two guides targeting exon 4 were designed using the http://crispr.mit.edu tool (guide 1: GAACCCACTATGGAAAGAGA; guide 2: GAGGCCCTCCCCATATGAGA) and were cloned into the pX458 plasmid (Addgene, 48138) as previously described[43]. The resulting plasmids were then used to transfect the cells as detailed above. Single transfected clones were expanded and assessed for Cas9-induced mutations. Genomic DNA was isolated by incubating cell pellets in 0.1 mg/ml of proteinase K (Sigma, 03115828001) in TE buffer at 50 °C for 2 h, followed by 5 min at 99 °C. The sequence flanking the guide-targeted sites was amplified from the genomic DNA by polymerase chain reaction (PCR) in a Biometra T3000 Thermocycler (30 s at 98 °C; 30 cycles of 10 s at 98 °C, 20 s at 58 °C, 20 s at 72 °C; and elongation for 7 min at 72 °C) using the Phusion High-Fidelity DNA Polymerase (NEB, M0530S) according to the manufacturer's instructions. Primers including Nextera overhangs were used (F- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTGCCCCTCCCATTTATGTA; R- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGTTCCAAGCCAGCATTTT), allowing library preparation with the Nextera XT Kit (Illumina, 15052163), and sequencing was performed using the Illumina MiSeq system according to the manufacturer's instructions. An ES cell clone showing a 77 base-pair deletion in exon 4 that inactivated *Tal1* gene expression was then injected into C57BL/6 E3.5 blastocysts. Chimeric embryos were subsequently transferred into recipient females at 0.5 days of pseudopregnancy following mating with vasectomized males, as described previously[44].

Chimeric embryos were collected at E8.5 (7 embryos), dissected, and single-cell suspensions were generated from pooled embryos as described above. Given the low detection rate of the tdTomato transcript (Extended Data Fig. 7b), single-cell suspensions were sorted into tdTomato$^+$ and tdTomato$^-$ samples using a BD Influx sorter with DAPI at 1 μg/ml (Sigma) as a viability stain for subsequent 10x scRNA-seq library preparation (v.2 Chemistry) and sequencing on an Illumina HiSeq 4000 platform. A total of 27,817 tdTomato$^-$ and 28,305 tdTomato$^+$ cells passed our quality control measures (before doublet and stripped nuclei removal, see below). Supplementary Table 5 contains metadata for each sequenced cell. Flow cytometry of chimeric embryos was performed in parallel using a BD Fortessa cytometer. Cells were stained with the conjugated antibodies CD45–APC–Cy7 (1:200; BD Pharmingen, 557659, clone 30-F11, 6126662), CD41–BV421 (1:200; Biolegend, 133911, clone MWReg30, B216311), Ter119–PerCP–Cy5 (1:200; Biolegend, 116227, clone TER-119, B169767), and CD71–FITC (1:400; BD Pharmingen, 553266, clone C2, 2307673), with Fc block CD16/32 (1:100; eBioscience, 14-0161-85, clone 93, 4316103), and DAPI at 1 μg/ml (Sigma) as a viability stain. For the wild-type into wild-type experiment, a parental tdTomato$^+$ *Tal1*$^{+/+}$ line was injected into C57BL/6 E3.5 blastocysts and processed as for the *Tal1*$^{-/-}$ samples. Three pooled embryos were used for scRNA-seq, and 1,077 tdTomato$^-$ and 2,454 tdTomato$^+$ cells passed quality control (Supplementary Table 6). Chimaera sample sizes were dependent on the number of viable embryos that did not show excessive global biases towards host or injected cells (that is, very low or high fluorescence). Two E9.5 embryos were individually analysed by flow cytometry as described above.

10x Genomics data pre-processing. Raw files were processed with Cell Ranger 2.1.1 using default mapping arguments. Reads were mapped to the mm10 genome and counted with GRCm38.92 annotation, including tdTomato sequence for chimaera cells. HTML reports that provide code, greater detail, and diagnostic plots for the following steps are available at https://github.com/MarioniLab/EmbryoTimecourse2018. Singularity containers are also available, providing direct access to the same software versions that were used in this analysis.

Swapped molecule removal. Molecule counts that were derived from barcode swapping were removed from all 10x samples by applying the DropletUtils

function 'swappedDrops' (default parameters) to groups of samples (where a sample is a single lane of a 10x Chromium chip) that were multiplexed for sequencing.

Cell calling. Cell barcodes that were associated with real cell transcriptomes were identified using emptyDrops[45], which assesses whether the RNA content associated with a cell barcode is significantly distinct from the ambient background RNA present within each sample. A minimum unique molecular identifier (UMI) threshold was set at 5,000, and cells with $P < 0.01$ (Benjamini–Hochberg-corrected) were considered for further analysis. The ambient RNA profile was determined from barcodes associated with fewer than 100 UMIs for the atlas and fewer than 60 UMIs for the chimaeras. We reproduced our analysis pipeline with a lower UMI threshold of 1,000 and found that no new cell types were present, justifying our rigorous threshold (Extended Data Fig. 2c).

Quality control. Cell libraries with low complexity (fewer than 1,000 expressed genes) were excluded. Cells with mitochondrial gene-expression fractions greater than 2.37%, 2.18%, and 3.35% for each of the wild-type atlas, *Tal1*$^{-/-}$ chimaeras, and wild-type chimaeras, respectively, were excluded. The thresholds were determined by considering a median-centred median absolute deviation (MAD)-variance normal distribution; cells with mitochondrial read fraction outside of the upper end of this distribution were excluded (where outside corresponds to $P < 0.05$; Benjamini–Hochberg-corrected).

Normalization. Transcriptome size factors were calculated for each dataset separately (atlas, *Tal1*$^{-/-}$ chimaeras, wild-type chimaeras), using 'computeSumFactors' from the scran R package[46]. Cells were pre-clustered with the 'quickCluster' function using the parameter 'method=igraph' (using the scran R package), and minimum and maximum cluster sizes of 100 and 3,000 cells, respectively. Raw counts for each cell were divided by their size factors, and the resulting normalized counts were used for further processing.

Selection of highly variable genes. Highly variable genes (HVGs) were calculated using 'trendVar' and 'decomposeVar' from the scran R package, with loess span of 0.05. Genes that had significantly higher variance than the fitted trend (Benjamini–Hochberg-corrected $P < 0.05$) were retained. Genes with mean $\log_2$ normalized count $<10^{-3}$; genes on the Y chromosome; the gene *Xist*; and the reads mapping to the tdTomato construct (where applicable) were excluded.

Batch correction. Batch effects were removed using the 'fastMNN' function in the scran R package on 50 principal components computed from the HVGs only. Correction was performed first between the samples of each time point, merging sequentially from the samples containing the most cells to the samples containing the least. Time points were then merged from oldest to youngest, and the mixed time point was merged between E7.25 and E7.0 (Extended Data Fig. 2d). This method was carried out on the whole atlas dataset, and recalculated separately on the subsets of cells considered in Figs. 2, 3 for their respective analyses. Euclidean distances calculated from this batch-corrected principal component analysis were used for all further analysis steps (for example, nearest-neighbour graphs).

Doublet removal. First, a doublet score was computed for each cell by applying the 'doubletCells' function (scran R package) to each 10x sample separately. This function returns the density of simulated doublets around each cell, normalized by the density of observed cell libraries. High scores indicate high doublet probability. We next identified clusters of cells in each sample by computing the first 50 principal components across all genes, building a shared nearest-neighbour graph (10 nearest neighbours; 'buildSNNGraph' function; scran R package), and applying the Louvain clustering algorithm ('cluster_louvain' function; igraph R package; default parameters) to it. Only HVGs (calculated separately for each sample) were used for the clustering. This procedure was repeated in each identified cluster to break the data into smaller clusters, ensuring that small regions of high doublet density were not clustered with large numbers of singlets. For each cluster, the median doublet score was considered as a summary of the scores of its cells, as clusters with a high median score were likely to contain mostly doublets. Doublet calls were made in each sample by considering a null distribution for the scores using a median-centred MAD-variance normal distribution, separately for each sample. The MAD estimate was calculated only on values above the median to avoid the effects of zero-truncation, as doublet scores cannot be less than zero. All cells in clusters with a median score at the extreme upper end of this distribution (Benjamini–Hochberg-corrected $P < 0.1$) were labelled as doublets. A final clustering step was performed across all samples together to identify cells that shared transcriptional profiles with called doublets, but escaped identification in their own samples. Clusters were defined using the same procedure as was applied to each sample, with the exceptions that sub-clustering was not performed, and batch-corrected principal components were used (see 'Batch correction', above). To identify clusters that contained more doublets than expected, we considered for each cluster the fraction of cell libraries that were called as doublets in their own samples. We modelled a null distribution for this fraction using a median-centred, MAD-estimated variance normal distribution as described for the median doublet score in each sample, above, and called doublets from the distribution as in each sample, above.

Stripped nucleus removal. Five of the clusters found in the across-sample clustering step above (see 'Doublet removal') contained cells with considerably lower mitochondrial gene expression and smaller total UMI counts compared with other clusters. We assumed that these clusters consisted of nuclei that had been stripped of their cytoplasm in the 10x droplets, and therefore excluded them from downstream analyses.

Density estimation. The density of cells in gene-expression space was calculated using a tricube kernel on the top 50 batch-corrected principal components. The median distance of all cells to their fiftieth nearest neighbour was used to define the maximum distance for the kernel.

Smart-seq2 data pre-processing. *Mapping.* Reads were mapped to the mm10 genome using GSNAP[47] (v.2015-09-29) with default arguments except for 'batch' = 5. HTSeq[48] was subsequently used to count the number of reads mapped to each gene, using GRCm38.92 for annotation.

Quality control. Three criteria were used to identify and discard poor-quality cells: (1) Number of mapped reads to nuclear genes <50,000; (2) number of genes detected <4,000; and (3) proportion of reads mapping to mitochondrial genes >10%. Cell libraries for which any of these criteria were met were discarded. Of the 288 cell libraries prepared, 250 passed our quality control.

Normalization. Cells were size-factor normalized as above (scran R package).

Visualization. UMAPs were calculated using Scanpy (v.1.2.2[49]; 'scanpy.api.tl.umap'). The 20 nearest neighbours in the batch-corrected principal component analysis were considered, with default parameters except for 'min.dist' = 0.7.

Force-directed graphs considered the 10 nearest neighbours of each cell in a 15-dimension diffusion space calculated on the first 50 principal components of the HVG-subset data (using the Scanpy v.1.2.2 function 'tl.diffmap' and Scanpy v.0.4.4[49] function 'utils.comp_distance'). Edges were unweighted, and the layouts were generated in Gephi (v.0.9.2)[50] using the ForceAtlas2 algorithm[51].

Endoderm diffusion maps were calculated from batch-corrected principal component coordinates, using the destiny R package with function 'DiffusionMap' (default settings).

Graph abstraction[52] was computed using the 'tl.aga' function from the Scanpy v.1.2.2 module in Python, and edges were drawn using the adjacency confidence matrix. For Extended Data Fig. 3e–f, graph abstraction was computed on the clusters annotated in Fig. 1c, and the threshold for connection of clusters was set to 0.23. For Fig. 3b, the clusters in Fig. 3a were used, and the threshold was set to 0.2, and for Extended Data Fig. 5a, thresholds of 0.95 (top panels) and 0.45 (bottom panels) were used. The generation of the clusters for the plots in Extended Data Fig. 5a is described below ('Endoderm analysis').

Clustering and cell annotation. A shared nearest-neighbour graph (considering the 10 neighbours of each cell) was constructed using the 50 batch-corrected principal components of the HVG-subset expression data and Euclidean distance ('buildSNNGraph', scran R package). Clusters were called from this graph using the Louvain algorithm ('cluster_louvain' with default parameters; igraph R package). To identify finer substructure from these top-level clusters, each cluster underwent a second round of clustering using the same method as above with the same batch-corrected principal component coordinates (that is, by subsetting from the batch-corrected coordinates). To assess the connectivity across all sub-clusters in the dataset, graph abstraction was then implemented on the first 15 diffusion components computed on the batch-corrected principal component analysis space (50 principal components). Within each top-level cluster, we considered distances between the sub-clusters based on their graph abstraction connectivity. Specifically, for a connectivity confidence score between clusters of x, we considered a distance of $1 - x$. Ward-linkage hierarchical clustering was performed to evaluate sub-cluster relatedness ('scipy.cluster.hierarchy' module; Python 3.4). Sub-clusters with distances of less than 1.6 were merged. The merged sub-clusters were then annotated by examination of marker gene expression. Sub-clusters without a unique identity according to marker gene expression were manually merged with their closest sub-cluster to form final cell-type annotations. Unannotated sub-clusters (that is, before merging) are available in Supplementary Table 4.

Stability of the atlas under downsampling. To test the stability of the atlas with regard to the size of the dataset, we sampled with replacement cell-type labels from the atlas dataset. We performed 50 samplings for each of the sizes of sample (1,000–116,312 cells) and calculated for each cell type the ratio of the standard deviation of cell-type label frequency to mean cell-type frequency. The ratios are shown in Extended Data Fig. 3d. Note that when the atlas was downsampled to less than half of its full size (50,000 cells), the standard deviation of cell-type frequency remained less than 10% of the mean for all cell types.

Endoderm analysis: cell selection and annotation of the gut. Cell types annotated as anterior primitive streak, definitive endoderm, visceral endoderm, and gut were selected for further analysis. A batch correction was computed for this cell set, using the same method as described above.

Cells with the gut cell-type label from the collection time points E8.25 and E8.5 were selected. We constructed a shared nearest-neighbour graph on their

batch-corrected principal component coordinates (that is, by subsetting from the coordinates from the endoderm-specific correction; 'buildSNNGraph' function; scran R package; 10 nearest neighbours), and clustered cells using the Louvain algorithm ('cluster_louvain' function; igraph R package; default parameters).

Endoderm analysis: gut tube pseudospatial ordering. E8.5 cells from the gut clusters (Fig. 2c) were selected and a diffusion map was constructed from their batch-corrected principal component coordinates. DPT was calculated for each cell starting from the pharyngeal endoderm cell with minimum value on diffusion component 2.

Differential expression analyses were performed using the 'findMarkers' function in the scran R package, using the 10x sample as a blocking factor. Significantly differentially expressed genes were considered as those with Benjamini–Hochberg-corrected $P < 0.1$. Hindgut differentially expressed genes were tested against an absolute fold change of 0.5.

Endoderm analysis: transport maps. This approach considers each cell as a unit of mass that can be 'transported' to other cells at consecutive time points[16]. By seeking to move these masses efficiently between time points (that is, minimizing the transcriptional differences between cells across which mass is moved) a mapping of expected descendant and ancestor cells can be identified. Importantly, this method allows the integration of both transcriptional and collection time-point information. Transport maps were constructed using the wot package in Python (v.0.2.1) using default settings except for skipping the dimension-reduction step, and instead using the batch-corrected principal components as input. In total, 100 randomly selected cells from each collection time point of ExE endoderm were added to the cells projected in Fig. 2c (Extended Data Fig. 4h). Cells from the mixed time points were excluded from the analysis.

Endoderm analysis: selecting cells for trajectories with the transport maps. For pushing mass forward through the graph (that is, identifying from which progenitor cells the gut clusters derived), we considered two starting populations—definitive endoderm (E7.0 cells labelled as anterior primitive streak or definitive endoderm) and visceral endoderm (E7.0 cells labelled as visceral endoderm). This stage was selected because each of the two populations still retained a very distinct transcriptional profile, and contained a large number of cells (>400 for each category). For pulling mass backward through the graph (that is, selecting cells for the definitive endoderm–hindgut 2 and visceral endoderm–hindgut 1 trajectories), we considered E8.5 cells from each of the gut clusters as terminal populations. For the cells in the visceral endoderm–hindgut 1 trajectory, we included all cells in which the largest mass contribution was to the hindgut 1 cluster. For cells in the definitive endoderm–hindgut 2 trajectory, we selected cells whose hindgut 2 mass contribution was greater than 90% of their largest mass contribution to any cluster. This approach allowed us to select cells that were committed to hindgut 2 (that is, with greatest mass towards hindgut 2), and also common progenitor cells, which show relatively balanced mass contributions to several terminal clusters. Cells with balanced mass contributions across clusters were not observed for the visceral endoderm–hindgut 1 trajectory, consistent with the hindward bias of the intercalated visceral endoderm cells.

Endoderm analysis: selecting cells for trajectories with graph abstraction. A shared 10-nearest-neighbour graph was constructed on the endoderm cell subset using the first 50 batch-corrected principal components of the HVG-subset expression data and Euclidean distance ('buildSNNGraph'; scran R package). Clusters were called from this graph using the Louvain algorithm ('cluster_louvain' with default parameters; igraph R package), and clusters that presented more substructure in the force-directed layout were further sub-clustered using the same pipeline. Cells were selected based on a manually curated parsimonious trajectory connecting hindgut 1 or 2 to the appropriate progenitor populations (Extended Data Fig. 5a, b).

Endoderm analysis: hindgut trajectories. Genes were clustered as in previous work[5], with some modifications. First, cells were ordered using DPT, which was calculated from a diffusion map built from the endoderm-specific batch-corrected principal component coordinates of the relevant subset of cells. DPT was calculated from a cell with the most extreme value on diffusion component 1, and direction along diffusion component 1 was selected to start from the youngest populations of cells. HVGs were calculated for each cell subset and only these genes were retained for subsequent clustering. For each retained gene, we fitted two ordinary least squares linear models (constant and degree-2 polynomial functions) that regressed the $\log_2$ normalized expression levels for each cell against the values of DPT calculated above. Genes for which the degree-2 polynomial fit the data better were retained (F test, Benjamini–Hochberg-corrected $P < 0.05$, R function 'anova.lm'). For each of these genes, we fitted a local regression to the expression level for each cell at their value of DPT (R function 'loess', span = 0.75). We then identified the predicted value of the loess fit for 1,000 uniformly spaced points across the DPT to provide smoothed gene expression estimates and avoid biasing clustering to regions of DPT with high cell density. The loess fits were scaled to a range of (0,1) to prevent clustering by expression level. The Pearson correlation distance between each gene was calculated as $([1 - x]/2)^{0.5}$, where x is the Pearson correlation, and hierarchical clustering was performed using the unweighted pair group method with arithmetic mean (UPGMA). The tree was cut with 'dynamicTreeCut' (R; minimum cluster size of 50 genes, otherwise default parameters).

Blood development analysis: cell clustering and differential expression. A 10-nearest-neighbour graph was constructed on the haemato-endothelial cell subset using the first 50 batch-corrected principal components of the HVG-subset expression data and Euclidean distance ('buildKNNGraph'; scran R package). Clusters were called from this graph using the Louvain algorithm ('cluster_louvain' with default parameters; igraph R package). Two clusters that presented higher substructure in the force-directed layout (one in EC, containing EC3–8; and one in the Haem/BP region, containing Haem3–4, BP3–4, My, and Mk) were further sub-clustered using the same pipeline but different k values: $k = 30$ for the EC cluster and $k = 15$ for the Haem/BP cluster.

Pairwise comparisons were performed using edgeR[53]. Dispersions were estimated using 'estimateCommonDisp' and 'estimateTagwiseDisp' tests, with 'exactTest' function and Benjamini Hochberg-corrected P values. All functions were used with default parameters.

Blood development analysis: mapping of Smart-seq2 data to the reference atlas. The Spearman correlation distance, $([1 - x]/2)^{0.5}$, where x is the Spearman correlation coefficient, was computed between each cell in the Smart-seq2 dataset and each cell in the endothelium cluster from the 10x atlas using the HVGs computed for the EC clusters of Fig. 3. The labels of the atlas endothelial cells were defined as the most frequent dissection location within the five nearest neighbours. If cells had an equal number of neighbours from two locations, they remained unassigned.

Blood development analysis: mapping of published embryonic blood dataset. To support the annotation of the myeloid cluster, atlas cells from Fig. 3a were mapped to a published dataset[54] containing haematopoietic cells collected between E9.5 and E11.5 (Extended Data Fig. 6g). The mapping was performed as for the Smart-seq2 dataset, using the HVGs computed for the published dataset. The data were processed as follows. The counts matrix with transcript counts per million (TPM) was downloaded from the GEO (accession GSE87038). Counts were log transformed as $\log_2(n/10 + 1)$, where n is the TPM value, and HVGs were calculated as in previous work[55]. Since cluster identities were not provided, the data were re-clustered using Louvain clustering on a k-nearest-neighbour graph with $k = 10$, considering only the HVGs. Clusters were subsequently merged to approximate the clusters and expression patterns of marker genes shown in figure 8 of the published study[54].

***Tal1*$^{-/-}$ chimaera analysis: mapping to the atlas.** To avoid mapping biases that derive from unequal numbers of atlas cells at each collection time point, each stage of the atlas was sub-sampled at random such that 10,000 cell libraries (that is, including doublets and stripped nuclei) were present at each time point. Cells from the mixed time point were excluded. Atlas stages E6.5 and E6.75 contained fewer cells (3,697 and 2,169, respectively) and were not downsampled; however, this cell number bias is likely to be unimportant as we would not expect cells from E8.5 chimaeras to map to these time points. We first constructed a 50-dimensional principal component space from the combined normalized log-counts of subsampled atlas cells (including doublets and stripped nuclei) and the cells from the samples that were to be mapped to the atlas. Batch correction was then performed on the atlas cells in the principal component space, as described above ('Batch correction'), to construct a single reference manifold for mapping. Samples to be mapped were then independently merged with the newly corrected atlas data (scran function 'fastMNN'), and the 10 nearest cells (Euclidean distance) in the atlas to each chimeric cell were recorded. Mapped time point and cell type of a given chimeric cell were defined as the mode of those of its nearest neighbours. Ties were broken by choosing the stage or cell type of the cell that had the lowest distance to the chimeric cell. Cells that mapped to doublet-labelled or stripped nucleus-labelled cells were excluded from downstream analyses. The robustness of this mapping was assessed by mapping one entire biological replicate of E8.0 cells onto the atlas, having removed the cells from the reference. 89.4% of these cells correctly mapped to their annotated cell type (Extended Data Fig. 7c), and 29.2% to the correct time point (Extended Data Fig. 7d; 83.1% of cells mapped to within one time point in either direction).

***Tal1*$^{-/-}$ chimaera analysis: visualization.** UMAP visualization of the data was performed as described above, using 50 batch-corrected principal components. Batches of cells of the same genotype were merged first, followed by merging across genotypes. To show the cell mapping with respect to atlas landscapes (for example, Fig. 4b), we coloured cells in the visualization by the closest atlas cell for each chimaera cell after mapping.

***Tal1*$^{-/-}$ chimaera analysis: remapping of cells to the haemato-endothelial landscape.** To ensure that we used the full resolution of our atlas, a subset of chimaera cells was mapped to the complete (that is, not subsampled) atlas dataset for the relevant cell types. The mapping procedure was repeated as described above, but for atlas and chimaera cells from the erythroid, haemato-endothelial,

blood progenitor, endothelial, and mixed mesoderm cell types. No downsampling was performed.

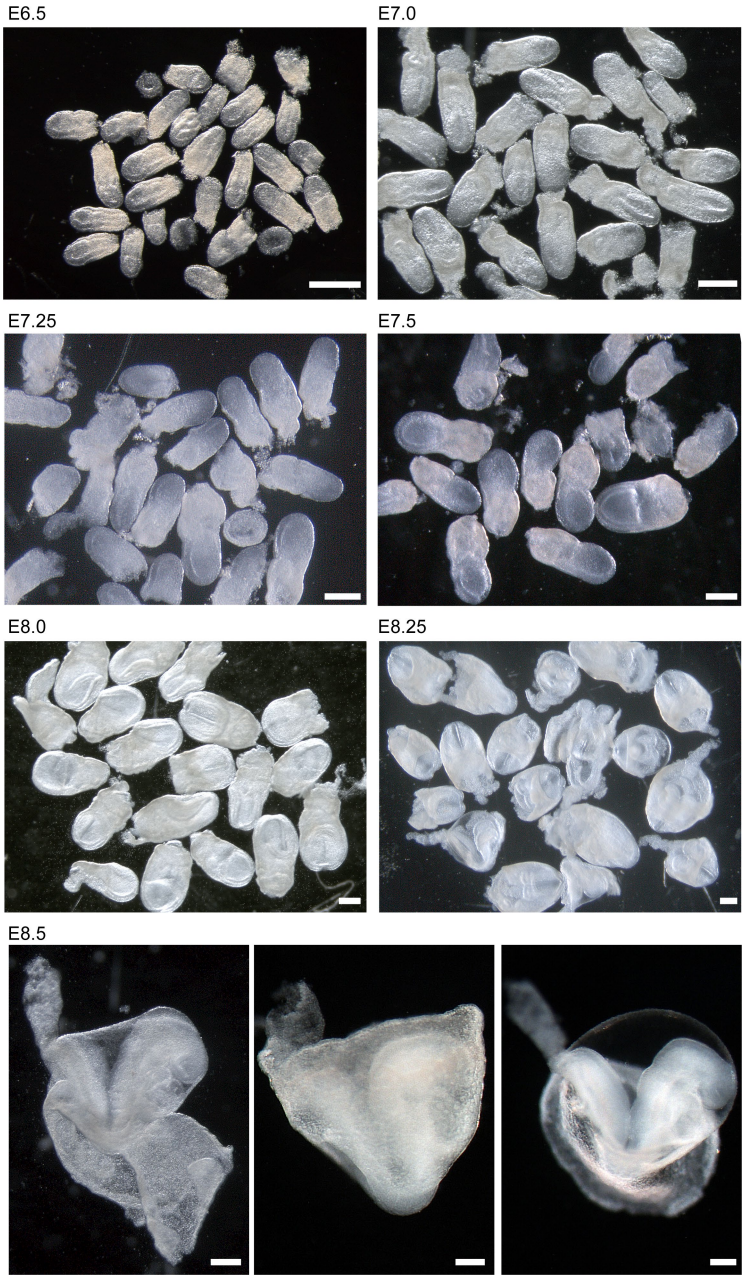
Relative contributions to atlas clusters from injected and host cells in the *Tal1*$^{-/-}$ chimaera. To compare the frequency of cells that contributed from each of the host and injected populations to atlas clusters, we first corrected for compositional differences between the populations. ExE endoderm, visceral endoderm, ExE ectoderm, parietal endoderm, blood progenitors 1–3, and erythroid 1–3 clusters were excluded from this analysis, as the injected (*Tal1*$^{-/-}$) cells cannot contribute to these lineages. The frequency of cells from each sub-cluster and the fold change ($\log_2$(fraction of injected cells mapped/fraction of host cells mapped) were then calculated (Fig. 4e). Owing to the absence of cells in the *Tal1*$^{-/-}$ samples, blood progenitor, and erythroid sub-clusters were not considered. Differential expression analyses were performed as in the 'Blood development analysis' sections.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

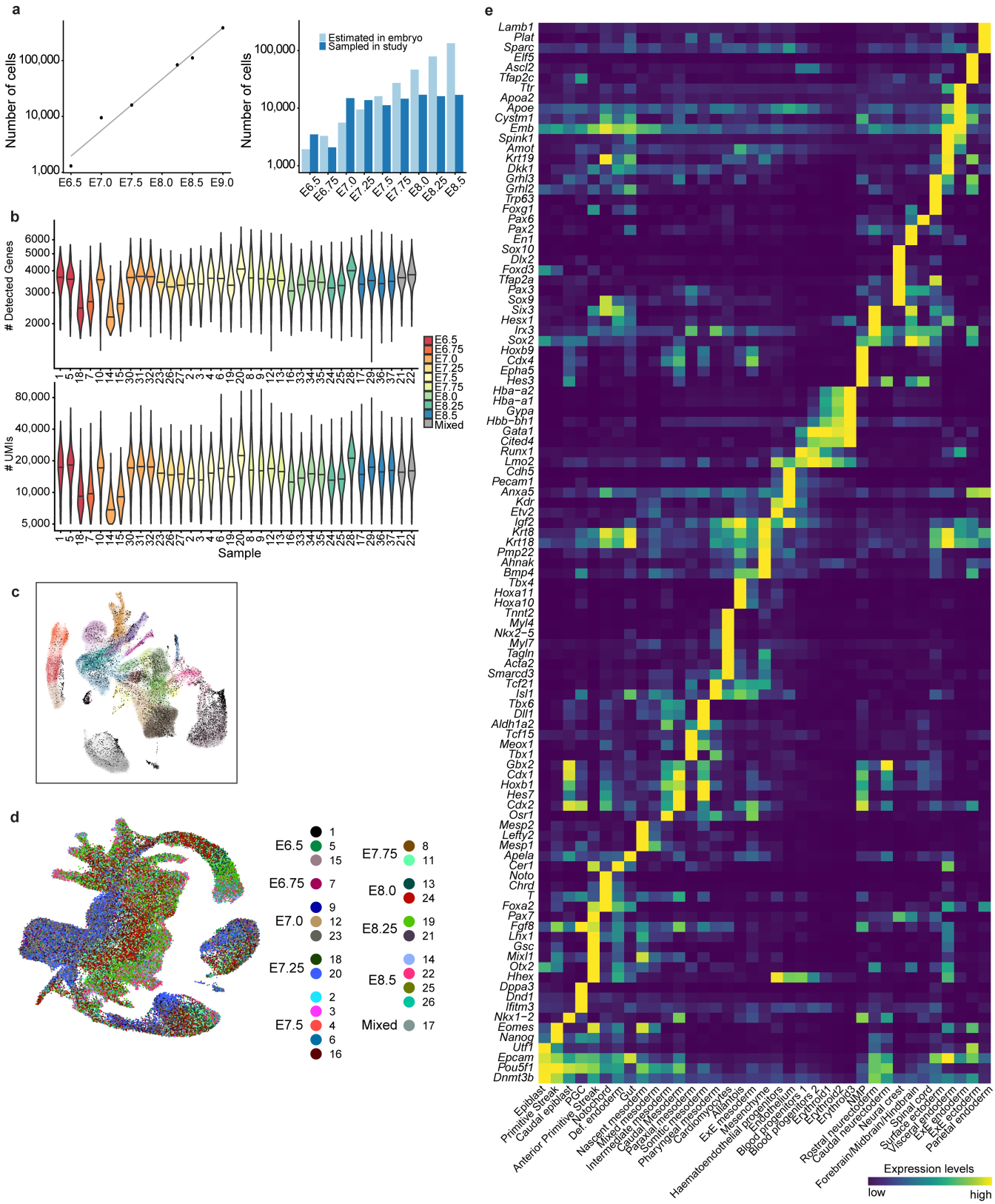
Data availability

Raw sequencing data are available on ArrayExpress with the following accessions: Atlas: E-MTAB-6967; Smart-seq2 endothelial cells: E-MTAB-6970; *Tal1*$^{-/-}$ chimaeras: E-MTAB-7325; wild-type chimaeras: E-MTAB-7324. Processed data may be downloaded following the instructions at https://github.com/MarioniLab/EmbryoTimecourse2018. Gene Expression Omnibus (GEO) accession GSE87038 was used to support the annotation of myeloid cells (see Methods). All code is available upon request, and at https://github.com/MarioniLab/EmbryoTimecourse2018. Our atlas can be explored at https://marionilab.cruk.cam.ac.uk/MouseGastrulation2018/. All other data are available from the corresponding authors on reasonable request.

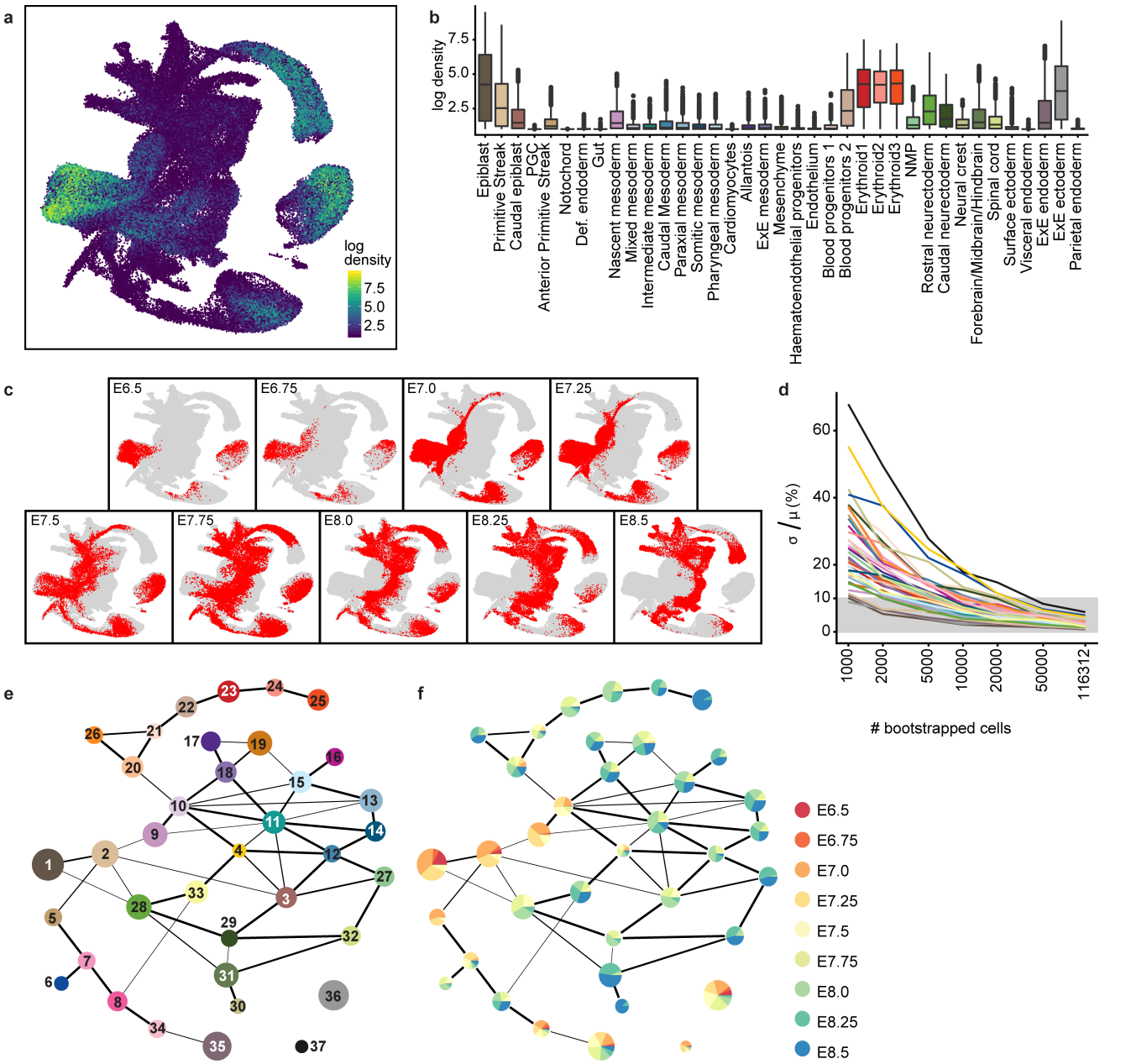
39. Srinivas, S. et al. Cre reporter strains produced by targeted insertion of *EYFP* and *ECFP* into the *ROSA26* locus. *BMC Dev. Biol.* **1**, 4 (2001).
40. Nichols, J. & Jones, K. Derivation of mouse embryonic stem (ES) cell lines using small-molecule inhibitors of Erk and Gsk3 signaling (2i). *Cold Spring Harb. Protoc.* **2017**, https://doi.org/10.1101/pdb.prot094086 (2017).
41. Ying, Q.-L. et al. The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).
42. Wray, J. et al. Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation. *Nat. Cell Biol.* **13**, 838–845 (2011).
43. Ran, F. A. et al. Genome engineering using the CRISPR-Cas9 system. *Nat. Protocols* **8**, 2281–2308 (2013).
44. Le Bin, G. C. et al. Oct4 is required for lineage priming in the developing inner cell mass of the mouse blastocyst. *Development* **141**, 1001–1010 (2014).
45. Lun, A. et al. Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Preprint at* https://www.bioRxiv.org/content/early/2018/04/04/234872 (2018).
46. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res.* **5**, 2122 (2016).
47. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
48. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
49. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
50. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. In *Third International AAAI Conference on Weblogs and Social Media* (AAAI, 2009).
51. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* **9**, e98679 (2014).
52. Wolf, F. A. et al. Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Preprint at* https://www.bioRxiv.org/content/early/2017/10/25/208819 (2017).
53. 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).
54. Dong, J. et al. Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis. *Genome Biol.* **19**, 31 (2018).
55. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
56. Kinder, S. J. et al. The orderly allocation of mesodermal cells to the extraembryonic structures and the anteroposterior axis during gastrulation of the mouse embryo. *Development* **126**, 4691–4701 (1999).



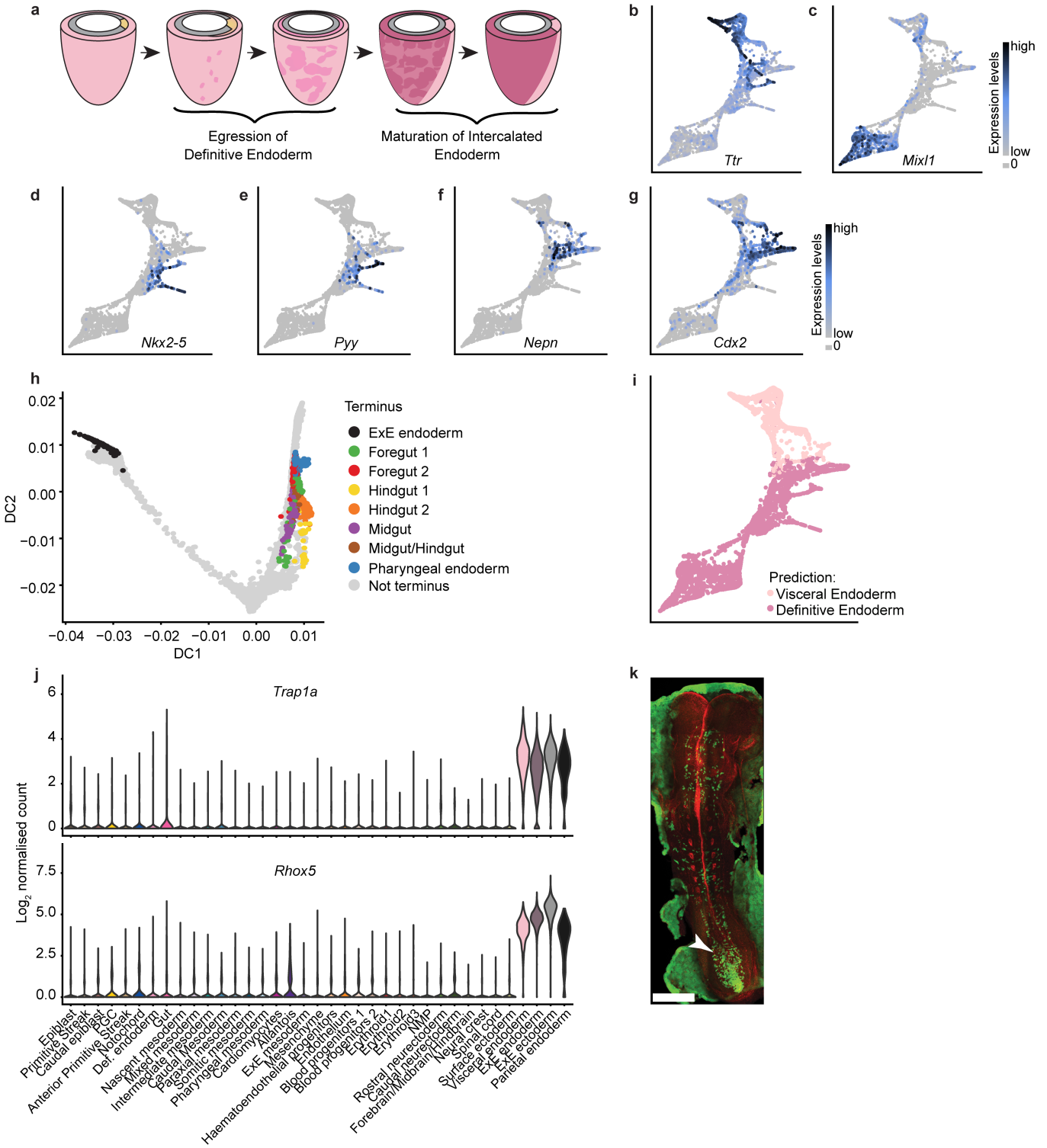
Extended Data Fig. 1 | Embryo images. Representative images of embryos collected at the time points indicated. Scale bars, 0.25 mm.



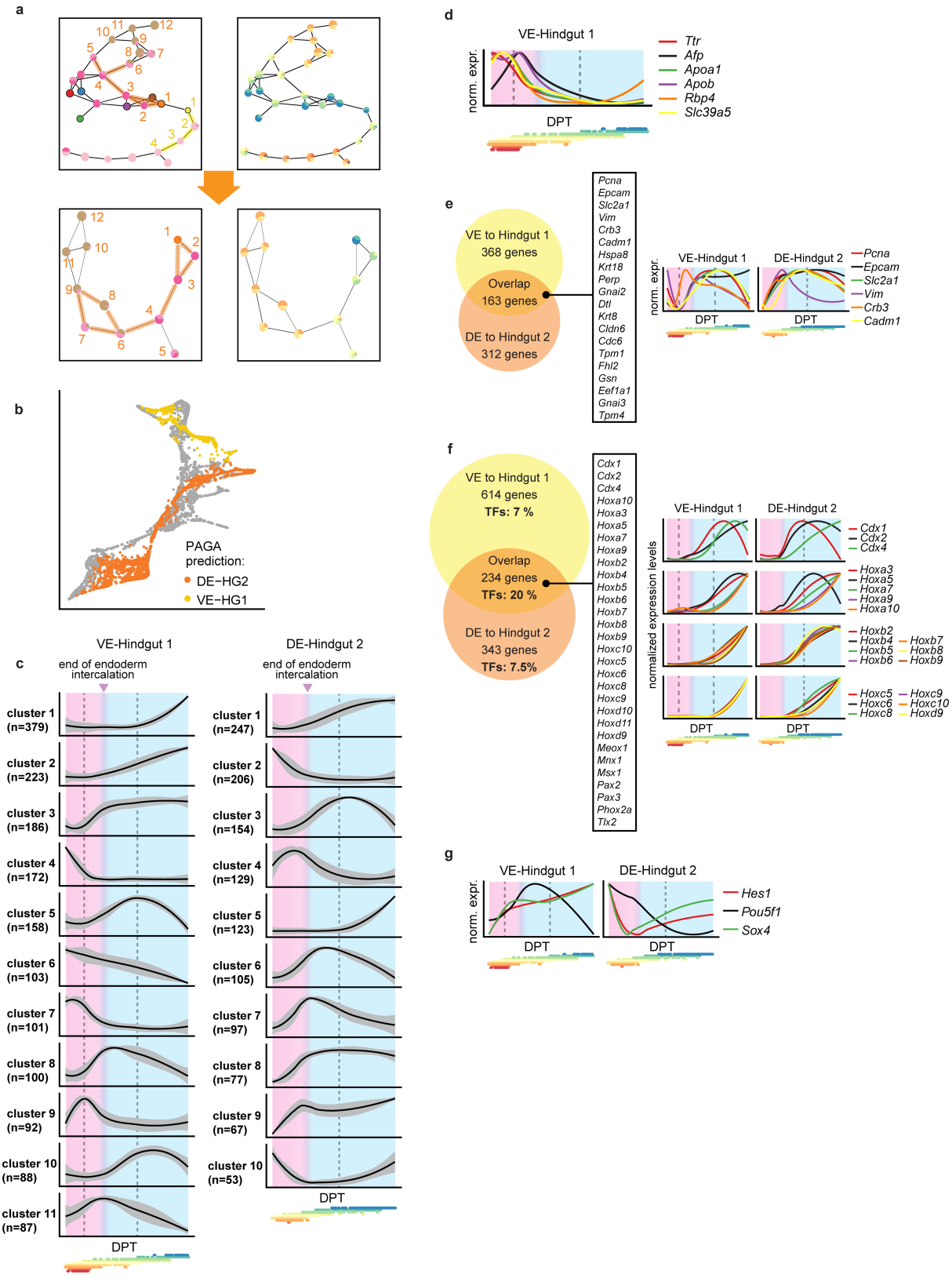
Extended Data Fig. 2 | Data quality control. a, Left, estimated number of cells present in a single mouse embryo at each time point. Points are values measured previously[24]; line is an ordinary least squares regression fit. Right, number of cells captured in this study compared with the estimated number of cells in the embryo from the left panel. **b**, Violin plots illustrating the number of detected genes (top) and unique molecular identifiers (UMIs; bottom) per cell per sample. Sample 11 failed quality control and is therefore not shown. Sample details are provided in Supplementary Table 1. **c**, UMAP highlighting additional cells identified when a reduced UMI threshold of 1,000 was considered. Additional cells are shown in black. Cells from the atlas are shown in the colour corresponding to their cell type (Fig. 1c). Note that all additional cells are present alongside cells from the atlas: no new cell types are found. **d**, UMAP plot as shown in Fig. 1c, with cells coloured by biological replicate, showing consistency between samples collected at the same time point. **e**, Heat map showing the mean gene expression of diagnostic markers (*y* axis) for each cell type (*x* axis). Genes are row-normalized. NMP, neuromesodermal progenitors; PGC, primordial germ cells.



Extended Data Fig. 3 | Lineage progression. a, UMAP plot as shown in Fig. 1c, coloured by the density of each cell in gene expression space; brighter coloured regions (towards yellow) are more densely sampled and darker regions (towards blue) are more sparsely sampled, relative to other regions in the atlas. Values shown are $\log_2(\text{density} + 1)$. **b**, Box plots summarizing the density per cell type. Values shown are $\log_2(\text{density} + 1)$. **c**, UMAP plots as shown in Fig. 1c, highlighting cells from each sampled time point and therefore illustrating the transcriptional progression along developmental time. **d**, Results of atlas stability testing (see Methods). *y* axis: ratio of the standard deviation of cell-type frequency to the mean cell-type frequency at different degrees of downsampling. Note that when the atlas is downsampled to less than half of its full size (50,000 cells), the standard deviation remains less than 10% of the mean for all cell types. x-axis is log-transformed. **e**, **f**, Abstracted graphs, which quantify the degree of similarity between the identified clusters to represent the underlying biological structure of the dataset. Nodes correspond to the annotated cell types, and edges reflect the confidence of adjacency between clusters (thicker edges indicate higher confidence). Node sizes increase as a function of the number of cells within each cluster. Nodes in **e** are coloured and numbered according to the legend shown in Fig. 1c. Nodes in **f** show the frequency of cells from each time point, excluding two samples of mixed time-point embryos.

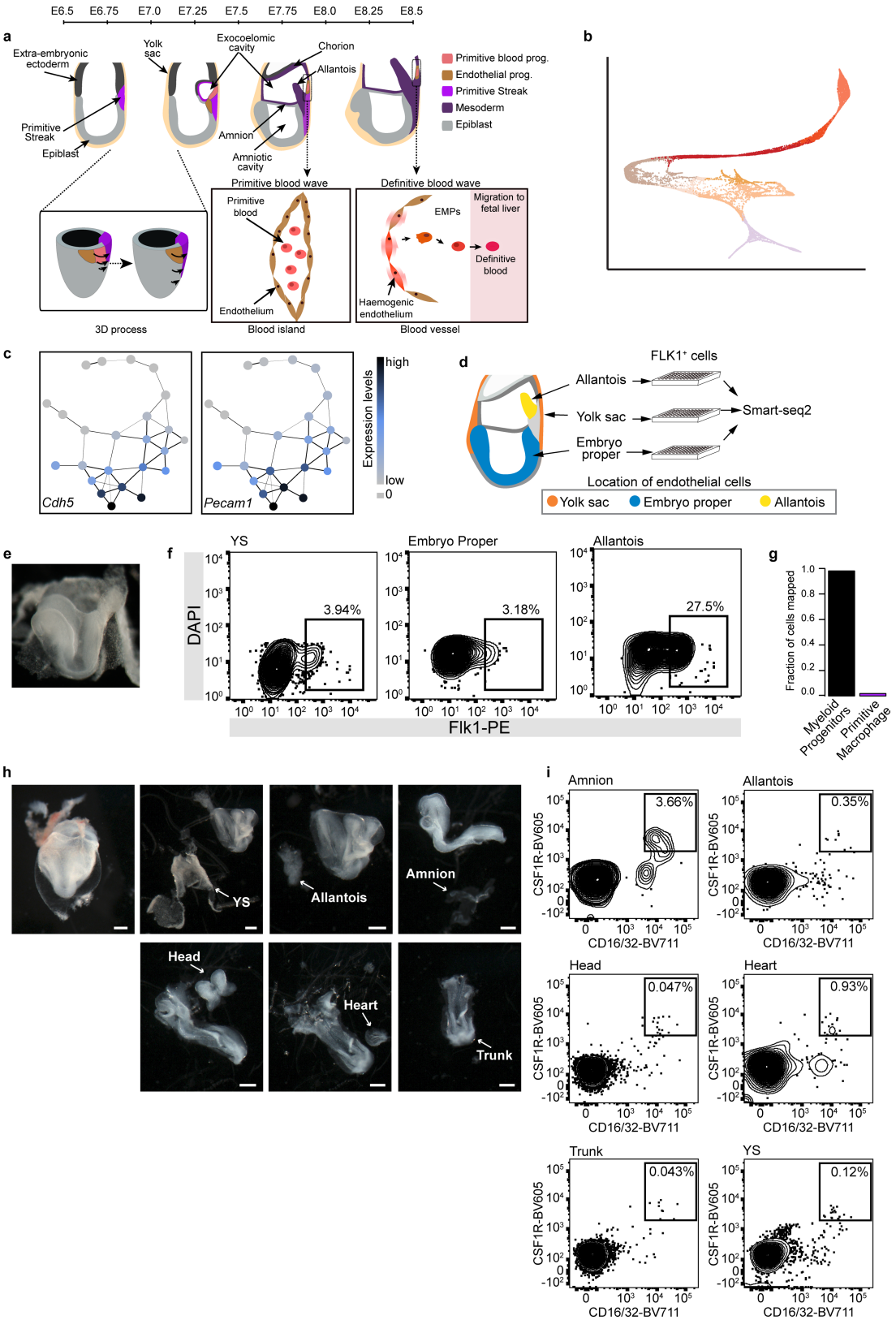


Extended Data Fig. 4 | Endoderm convergence. a, Schematic representing the process of definitive endoderm intercalation following gastrulation, and subsequent gut maturation. Adapted from a previous publication[9]. **b**–**g**, Gene expression levels of *Ttr* (**b**), *Mixl1* (**c**), *Nkx2-5* (**d**), *Pyy* (**e**), *Nepn* (**f**), and *Cdx2* (**g**), overlaid on the Fig. 2a force-directed graph. **h**, Diffusion map of cells selected for transport map construction; cells selected as termini for pulling mass backward through the transport maps are coloured. **i**, Results of pushing mass forward through the transport maps are shown on the force-directed layout. **j**, Violin plots showing expression levels of *Trap1a* and *Rhox5* in all cell-types of the full atlas. **k**, Dorsal view of a whole-mount fluorescence image of a *Ttr::cre*; R26R::YFP embryo at E8.5. Green, YFP; red, phalloidin. Arrowhead denotes the increased Ttr–YFP staining in the posterior region of the gut. Scale bar, 300 μm.



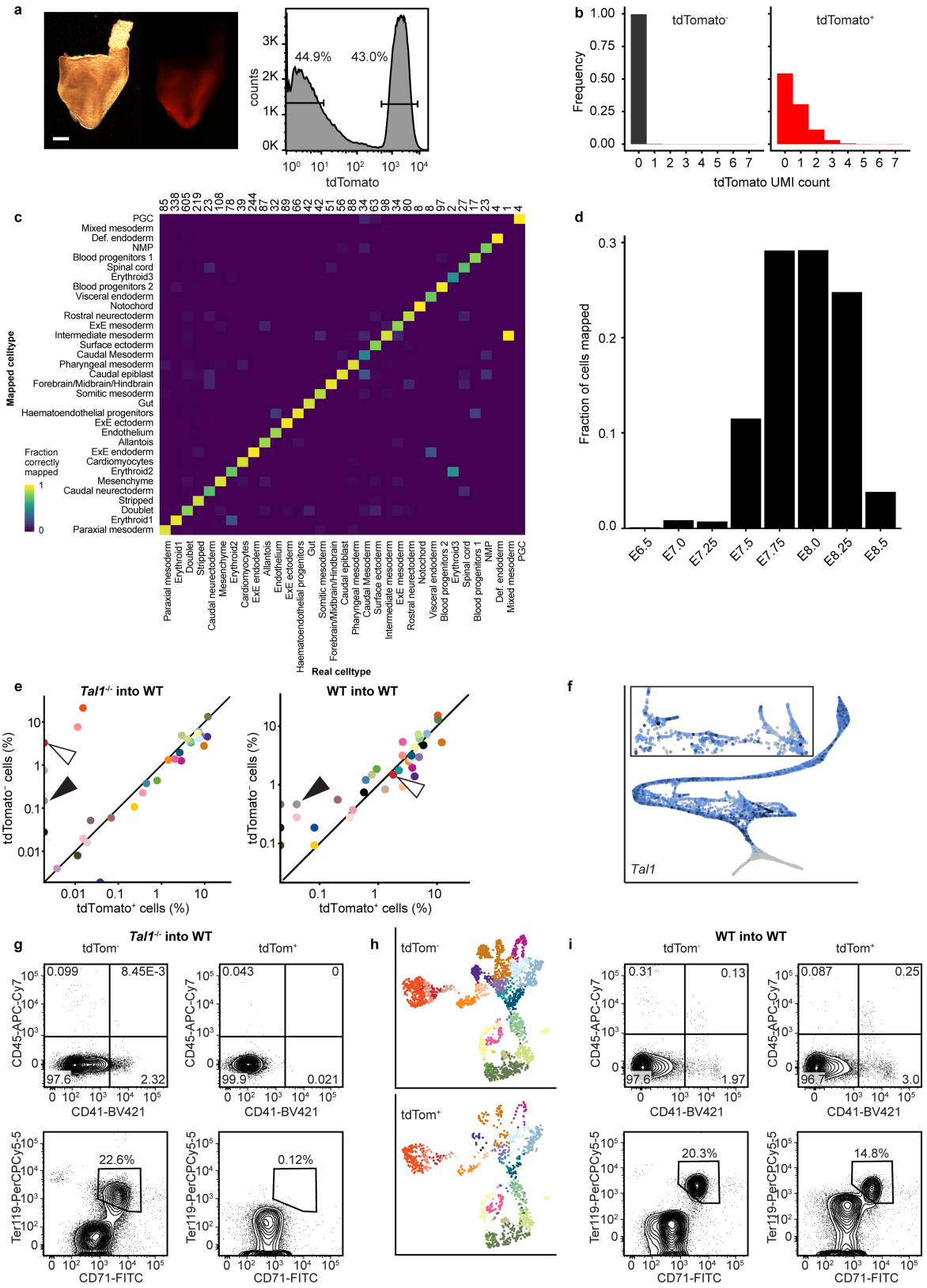
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Endoderm trajectories. a, Top, graph abstraction of the endoderm landscape after fine sub-clustering as an alternative method to resolve which cells should be part of the visceral endoderm (VE)–hindgut 1 trajectory or the definitive endoderm (DE)–hindgut 2 trajectory (supporting transport maps; see Methods). Edges along the VE–hindgut 1 trajectory are highlighted in yellow (nodes 1–4; yellow numbers). Edges along the DE–hindgut 2 trajectory are highlighted in orange (nodes 1–12; orange numbers). Bottom, graph abstraction with the subset of nodes related to the DE–hindgut 2 trajectory to resolve the origin of cluster 4 (between 5 and 6 in the top panel). Resulting DE–hindgut 2 trajectory includes clusters 1–4 and 6–9. The right-hand panels overlay information about the composition of each cluster by developmental stage. **b**, Force-directed graph coloured by partition-based graph abstraction (PAGA) trajectories. Note that this independent approach for trajectory identification reaches very similar results to those inferred by the transport maps in Fig. 2h. HG1, hindgut 1; HG2, hindgut 2. **c**, Gene-normalized dynamics of all clusters found along the VE–hindgut 1 and the DE–hindgut 2 trajectories (*x* axis: DPT along the trajectory; *y* axis: normalized expression levels). The black line is the mean fitted expression level across all genes in each cluster; the grey shading indicates the standard deviation along the trend across all genes in the cluster; the pink area highlights intercalation process; and the blue area highlights gut maturation steps. Vertical dashed lines correspond to additional stages in the process, deduced from the changes in gene expression trends. Points below the plots are the DPT coordinates of cells from each time point coloured according to time point as in Fig. 2b (from E6.5 in red to E8.5 in blue). **d**, Gene-normalized dynamics of VE genes along the VE–hindgut 1 trajectory, indicating VE maturation before the intercalation stage. Plot design is as in **c**. **e**, Left, Venn diagram of genes that were upregulated during the intercalation process in both VE–hindgut 1 (in clusters 3, 5, 8, and 11) and DE–hindgut 2 (in clusters 4, 6, 7, 8, and 9) trajectories. The overlapping fraction was enriched in genes that are a signature of epithelial remodelling (top 20 genes are listed). Right, gene-normalized dynamics of illustrative genes (*Pcna*, *Epcam*, *Slc2a1*, *Vim*, *Crb3*, and *Cadm1*) along the trajectories. **f**, Left, Venn diagram of genes that were upregulated after the intercalation process in both trajectories (VE–hindgut 1: clusters 1, 2, 5, and 10; DE–hindgut 2: clusters 1, 3, 5, and 10). The overlapping fraction was enriched in genes that encode transcription factors (TFs), including a large subset of homeodomain proteins (genes are listed). Right, gene-normalized dynamics of *Hox* and *Cdx* genes along the trajectories. **g**, Gene-normalized dynamics of transcription factors that were upregulated specifically in the VE–hindgut 1 trajectory during endoderm intercalation. Points below the *x* axis in **d**–**g** are as in **c**.



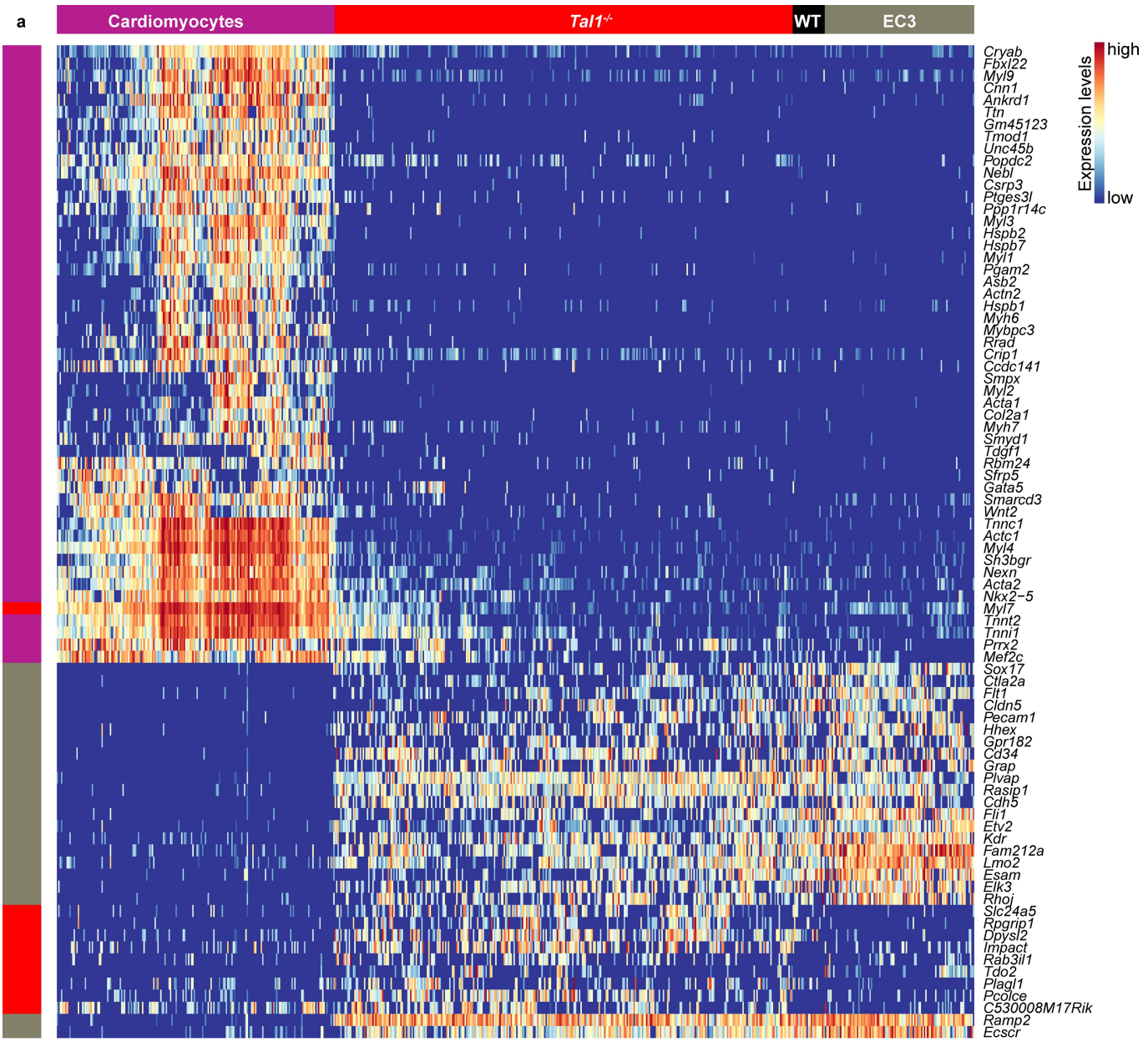
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Blood development. a, Diagram illustrating the two waves of embryonic blood development. At E6.5, gastrulation begins. Previous work using transplantation assays has shown that the proximo-posterior epiblast cells closest to the primitive streak at this stage (red) mainly give rise to primitive erythroid cells in the yolk sac, whereas the epiblast cells located in the middle of the embryo at E6.5 but closer to the primitive streak at a later stage are enriched for endothelial progenitors[56]. At E7.5, blood islands are apparent (zoomed box of primitive blood wave), where primitive erythroid cells are surrounded by endothelium. At around E8.25, some endothelial cells (haemogenic endothelium) undergo an endothelial-to-haematopoietic transition and become EMPs, which migrate to the fetal liver and give rise to definitive erythrocytes. Adapted from a previous study[38]. **b**, Force directed layout of Fig. 3a coloured by original clusters from Fig. 1c. **c**, Gene expression levels of *Cdh5* and *Pecam1*, overlaid on the graph abstraction visualization from Fig. 3b. **d**, Experimental design to isolate FLK1$^+$ cells from yolk sac, allantois, and embryo proper for Smart-seq2 scRNA-seq. **e**, Representative image of an embryo collected for the transcriptional analysis of endothelial cells from the yolk sac, allantois, and embryo proper. **f**, Sorting strategy of FLK1$^+$ cells from the yolk sac, embryo proper, and allantois on live cells (DAPI). *x* axis: FLK1 intensity. *y* axis: DAPI intensity. **g**, Evidence to support myeloid annotation of the myeloid cell cluster in Fig. 3. Haemato-endothelial cells from Fig. 3a were mapped to a published dataset[54] that profiled haematopoietic cells collected at E9.5, E10.5, and E11.5 from different organs. Bar charts show the fraction of atlas cells in the myeloid cell cluster mapped to the clusters defined in figure 8 of the previous study[54]. **h**, Representative images of the dissected regions collected to study the location of CSF1R$^+$CD16/32$^+$ cells. Scale bars, 0.25 mm. **i**, Flow cytometry plots indicating the frequency of CSF1R$^+$CD16/32$^+$ cells in each embryonic region. Two biological replicates were performed for this experiment: with pools of 12 and 13 embryos, respectively. Plots illustrate one biological replicate.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Analysis of *Tal1*$^{-/-}$ chimaeras. a, Left, representative chimaera embryo obtained at E8.5 (left: brightfield; right: tdTomato fluorescence; scale bar, 0.25 mm). Right, flow cytometry plot with tdTomato fluorescence distribution and sorting gates. **b**, Histograms showing the UMI counts for the tdTomato construct in both tdTomato^{-} and tdTomato^{+} fractions in the *Tal1*$^{-/-}$ into wild-type experiment (see Methods). **c**, **d**, Control mapping results of an E8.0 biological replicate that was removed and mapped back to the atlas. **c**, Heat map showing the fraction of cells of each labelled cell type that mapped to each cell type in the reference atlas. Numbers above columns indicate the number of cells in each category. Of these cells, 89.4% correctly mapped to their annotated cell type. **d**, Histogram showing the fraction of cells from the E8.0 biological replicate that mapped to each time point in the reference: 29.2% of cells mapped to the correct time point, and 83.1% of cells mapped within one time point (that is, 6 hours) in either direction. **e**, Scatter plot comparing the percentage of tdTomato^{+} cells against tdTomato^{-} for each cell type in both *Tal1*$^{-/-}$ into wild-type (WT; left) and wild-type into wild-type (right) experiments. Black arrowheads indicate extra-embryonic tissues; white arrowheads indicate haematopoietic tissues. **f**, Force-directed graph of blood-related lineages from the atlas (Fig. 3), coloured by *Tal1* expression levels. Darker colouring shows higher expression. **g**, Flow cytometry analysis of E8.5 *Tal1*$^{-/-}$ into wild-type chimaeras, showing the complete depletion of the haematopoietic markers CD41 and CD45 (top panels), as well as of the CD71^{+} Ter119^{+} erythroid fraction (bottom panels) in *Tal1*$^{-/-}$ tdTomato^{+} cells (right panels). **h**, UMAP plots of wild-type into wild-type experiment, showing balanced contributions to all embryonic lineages. **i**, Flow cytometry analysis of wild-type into wild-type chimaeras, showing balanced contributions to the haematopoietic lineage from both tdTomato^{+} and tdTomato^{-} cells at E9.5 (representative of 2 individual embryos).



Extended Data Fig. 8 | Transcriptional effects of disruption caused by *Tal1*. **a**, Heat map illustrating the row-normalized expression of genes that were upregulated in EC3-mapped *Tal1*$^{-/-}$ cells when compared with their closest neighbours in the atlas (labelled 'EC3') and EC3-mapped wild-type chimaera cells (labelled 'WT'). Genes *Gm45123* and *Fam212a* are also known as *5430431A17Rik* and *Inka1*, respectively. **b**, UMAP plots as in Fig. 1c, showing the expression of *Tdo2*, *Plagl1*, and *Pcolce*.

nature research

Corresponding author(s): Berthold Gottgens and John C. Marioni

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐ Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Cells were sequenced using HiSeq2500 and corresponding Illumina commercial software. FACS and flow cytometry data were collected using the BD FACSDIVA software.
Data analysis	Cell Ranger v2.1.1, GSNAP, HTSeq, Gephi v0.9.2. R packages: DropletUtils, emptyDrops, Matrix, scater, scran, igraph, Rtsne, irlba, edgeR,dynamicTreeCut, destiny, org.Mm.eg.db (v 3.4.1), GO.db (v 3.4.1), topGO (v 2.28.0) and GOstats (v2.42.0). Python modules: scanpy, scipy.cluster.hierarchy

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequencing data is available on Arrayexpress: Atlas – E-MTAB-6967; Smart-seq2 endothelial cells – E-MTAB-6970; Tal1 chimeras – E-MTAB-7325; WT chimeras – E-MTAB-7324. Processed data may be downloaded following the instructions at https://github.com/MarioniLab/EmbryoTimecourse2018. All code is available upon request, and at https://github.com/MarioniLab/EmbryoTimecourse2018. Our atlas can be explored at https://marionilab.cruk.cam.ac.uk/MouseGastrulation2018/. GEO accession GSE87038 was used to support the annotation of myeloid cells (see Methods).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Atlas sample sizes were chosen to maximise the number of recovered cells from each experiment and to obtain total cell numbers similar to the estimated cell numbers in mouse embryos at their respective stages. The sample sizes were also dependent on the number of viable embryos from each litter. Cells were partitioned to prevent overloading of a single 10X lane. Chimera sample sizes were dependent on the number of viable embryos that did not show excessive global biases towards host or injected cells (i.e. very low or high fluorescence). YS/EP/AL experiment: sample sizes were chosen based on the amount of viable endothelial cells recovered from the experiment and we aimed to have an equal (or very similar) number of endothelial cells from each of the dissected regions that was large enough (i.e. 96 per sample) to infer correlations with the atlas dataset.
Data exclusions	10X Sample 11 was excluded from analyses. The library generated from this sample was small; the sequencing saturation was very high (>90% vs. ~40% for other samples); and very few cells were called. It is therefore likely that this sample failed in some severe way during library generation and thus all cells were excluded. Per-cell QC metrics are available in the methods.
Replication	Multiple biological replicates (i.e. embryo pools) were generated for every timepoint of the atlas, except for E6.75.
Randomization	N/A - There was no case/control design
Blinding	N/A - There was no case/control design

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Flk1-PE antibody (1:100; eBioscience cat# 12-5821-83, clone Avas12a1, lot# E01819-1631); CD16/32-BV711 (1:200; Biolegend, cat# 101337, clone 93, lot# B251800); Fc block CD16/32 (1:100; eBioscience, cat# 14-0161-85, clone 93, lot# E03558-1640); CSF1R-BV605 (1:800; Biolegend; cat# 135517, clone AFS98, lot# B196541); 7AAD (1:200; BD Pharmigen; cat# 51-68981E, lot#

7061885); CD45-APC-Cy7 (1:200; BD Pharmingen, cat# 557659, clone 30-F11, lot# 6126662); CD41-BV421 (1:200; Biolegend, cat# 133911, clone MWReg30, lot# B216311), Ter119-PerCP-Cy5 (1:200; Biolegend, cat# 116227, clone TER-119, lot# B169767); CD71-FITC (1:400; BD Pharmingen, cat# 553266, clone C2, lot# 2307673); Fc block CD16/32 (1:100; eBioscience, cat# 14-0161-85, clone 93, lot# 4316103); Polyclonal Chicken Anti-GFP (1:100; Abcam, #ab13970, Lot No. GR3190550-2); Polyclonal Goat Anti-Chicken 488 (1:100; Invitrogen; #A11039, Lot No. 1899514); Phalloidin-Atto 555, Sigma (#19083)

Validation

Polyclonal Goat Anti-Chicken 488 (1:100; Invitrogen; #A11039, Lot No. 1899514) was validated by doing stainings on a litter which had both positive and negative embryos for the antibody. Secondary antibody only controls were also performed. Flow cytometry antibodies have been tested by the company. The descriptions provided here are taken from the company's websites: The Avas12a1 antibody has been reported for use in flow cytometric analysis and tested by flow cytometric analysis of bEnd.3 cells. Each lot of the Biolegend antibody clones was tested by flow cytometric analysis as follows: 93 and CSF-1R were tested on thioglycolate-elicited C57BL/6 mouse peritoneal macrophages, MWReg30 was tested on C57BL/6 platelets, TER-119 was tested on C57BL/6 mouse bone marrow cells. The the 30-F11 clone was tested by flow cytometric analysis of mouse splenocytes. The C2 clone was tested using BALB/c bone-marrow leukocytes. Additionally, unstained samples, Fluorescence Minus One (FMO's) and an isotype control for CSF1-R have been used during the experiments as a negative control.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

Mouse ESCs were derived from E3.5 blastocysts

Authentication

mESCs were derived from E3.5 blastocysts, mutant line was validated by sequencing the Cas9-targeted genomic region

Mycoplasma contamination

All lines were tested negative

Commonly misidentified lines
(See ICLAC register)

N/A

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

Mus musculus C57BL/6 males and females at E6.5, E6.75, E7.0, E7.25, E7.5, E7.75, E8.0, E8.25, E8.5, E9.5

Wild animals

The study did not involve wild animals.

Field-collected samples

The study did not involve samples collected from the field.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Yolk sac, allantois and embryo proper were dissected and placed into separate tubes. Single-cell suspensions were prepared by incubating the embryos with TrypLE Express dissociation reagent (Life Technologies) at 37 °C for 7 min and quenching with heat inactivated serum. Single cells were subsequently stained with Flk1-PE antibody (1:100; Biolegend) and DAPI as viability stain (1µg/ml; Sigma). Live Flk1+ cells were isolated by fluorescence-activated cell sorting (FACS) using a BD Influx sorter
Chimeric embryos were harvested at E8.5 of development, dissected, and single-cell suspensions were generated as described above . Single-cell suspensions were sorted into tdTomato+ and tdTomato- samples using a BD Influx sorter with DAPI at 1µg/ml (Sigma) as a viability stain.

Instrument

BD Influx sorter

Software

FACS DIVA

Cell population abundance

YS/AL/EP experiment: Yolk sac had 3.94% cells Flk1+, allantois had 27.5% cells Flk1+ and embryo proper had 3.72% cells Flk1+.
Experiment Chimeras: sorted tdTomato fractions were in the range of 45%.

Gating strategy

The first gate on FSC/SSC was used to exclude debris. Doublets were removed using the "Trigger Pulse Width". Positive and

Gating strategy

negative cells were defined using the unstained sample as reference for negative events and single-stained samples as reference for positive events.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.