

The emergent landscape of the mouse gut endoderm at single-cell resolution

Sonja Nowotschin^{1,6}, Manu Setty^{2,6}, Ying-Yi Kuo¹, Vincent Liu², Vidur Garg¹, Roshan Sharma², Claire S. Simon¹, Nestor Saiz¹, Rui Gardner³, Stéphane C. Boutet⁴, Deanna M. Church⁴, Pamela A. Hoodless⁵, Anna-Katerina Hadjantonakis^{1*} & Dana Pe'er^{2*}

Here we delineate the ontogeny of the mammalian endoderm by generating 112,217 single-cell transcriptomes, which represent all endoderm populations within the mouse embryo until midgestation. We use graph-based approaches to model differentiating cells, which provides a spatio-temporal characterization of developmental trajectories and defines the transcriptional architecture that accompanies the emergence of the first (primitive or extra-embryonic) endodermal population and its sister pluripotent (embryonic) epiblast lineage. We uncover a relationship between descendants of these two lineages, in which epiblast cells differentiate into endoderm at two distinct time points—before and during gastrulation. Trajectories of endoderm cells were mapped as they acquired embryonic versus extra-embryonic fates and as they spatially converged within the nascent gut endoderm, which revealed these cells to be globally similar but retain aspects of their lineage history. We observed the regionalized identity of cells along the anterior–posterior axis of the emergent gut tube, which reflects their embryonic or extra-embryonic origin, and the coordinated patterning of these cells into organ-specific territories.

The gut endoderm is the precursor of the respiratory and digestive tracts, and their associated organs^{1,2} (Fig. 1). Endoderm cells emerge twice during mammalian development. Primitive endoderm (PrE) cells arise at the blastocyst stage (around mouse embryonic day (E)3.5–4.0³) and predominantly contribute to parietal and visceral yolk sac endoderm. Later, at around E7.0, definitive (that is, embryonic) endoderm is specified from the pluripotent epiblast (EPI) at gastrulation⁴. Previous studies have revealed that the gut endoderm comprises cells of both PrE and definitive endoderm^{5–7} (Extended Data Fig. 1a). Common endodermal genes are expressed by both cell types, which hampers the marker-based discrimination of embryonic and extra-embryonic descendants^{8,9}. We therefore sought to characterize the transcriptional profiles of all endoderm populations within the mouse embryo from the blastocyst to midgestation (E3.5–E8.75), at which point the gut tube becomes regionally patterned along its anterior–posterior axis.

To analyse our data, we developed Harmony, an algorithm to bridge time points, and combined it with Palantir^{10,11}; we used these algorithms to construct a spatio-temporal map of the developing endoderm. Palantir infers cell fate potential, which provides a quantitative metric of plasticity and thus allows us to identify when fate decisions occur. The algorithm identified key bifurcation and convergence points of embryonic and extra-embryonic tissues that led to the establishment of distinct territories along the anterior–posterior axis of the gut tube at E8.75, before the overt appearance of endodermal organs. In sum, this study provides a comprehensive transcriptional characterization of the ontogeny of the endodermal organ system in a mammalian model.

Results

Cells were isolated from sequentially staged wild-type mouse embryos between E3.5 and E8.75 for single-cell RNA sequencing (scRNA-seq) (Fig. 1a). Owing to their small size, whole embryos were used for isolations at pre- and early post-implantation time points (E3.5–E5.5)

whereas endodermal tissues were isolated for cell-type enrichment from embryos between E6.5 and E8.75 (Extended Data Fig. 1b). To demarcate extra-embryonic (primitive and visceral) endoderm cells in the gut tube^{5,7}, we used the visceral endoderm-specific *Afp-GFP* mouse line¹² and isolated GFP-positive (extra-embryonic) and GFP-negative (embryonic) populations by flow cytometry, after tissue dissociation at E7.5–E8.75 (Fig. 1a, Extended Data Fig. 1c, Supplementary Fig. 1). We profiled 13 tissue types that were each collected in duplicate or triplicate, representing 112,217 cells in total (Fig. 1b, Extended Data Fig. 1d, Supplementary Fig. 2). We ran each sample through our processing pipeline^{10,11,13} (Extended Data Fig. 2a, Supplementary Note 1) and verified replicate reproducibility (Supplementary Fig. 2) before combining. Phenograph clustering¹⁴ identified cell types; labels were assigned on the basis of gene expression, and visualized using *t*-distributed stochastic neighbour embedding (*t*-SNE)¹⁵. Comparison to bulk RNA-seq data demonstrated that isolation and dissociation did not alter cell proportions or transcriptional profiles (Extended Data Figs. 1c, 2b).

Following recent successes in reconstructing developmental trajectories from scRNA-seq^{16–19}, we organized cells along trajectories to elucidate when and how fate decisions occur. We used Harmony to connect across time points (Extended Data Fig. 3, Supplementary Note 2). Asynchronous differentiation results in a subset of more mature cells at one time point being relatively closer to a subset of less mature cells in the following time point, which results in mutually similar cells across time points. Harmony uses these mutual nearest neighbours to construct an augmented *k* nearest neighbour (*k*NN) graph that connects time points (Extended Data Fig. 3a–f) without altering the underlying data matrix. This augmented graph can be used as input into any algorithms based on *k*NN graphs (Extended Data Fig. 3g).

We combined Harmony with Palantir^{10,11} (Supplementary Note 7), which takes as input a user-defined early ‘start cell’ and infers pseudo-time, and branch probabilities: denoting for each cell state, its probability to reach each of the terminal fates in the system. We define

¹Developmental Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ²Computational & Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ³Flow Cytometry Core Facility, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ⁴10x Genomics, Pleasanton, CA, USA. ⁵Terry Fox Laboratory, BC Cancer, Vancouver, British Columbia, Canada. ⁶These authors contributed equally: Sonja Nowotschin, Manu Setty. *e-mail: hadj@mskcc.org; peerd@mskcc.org

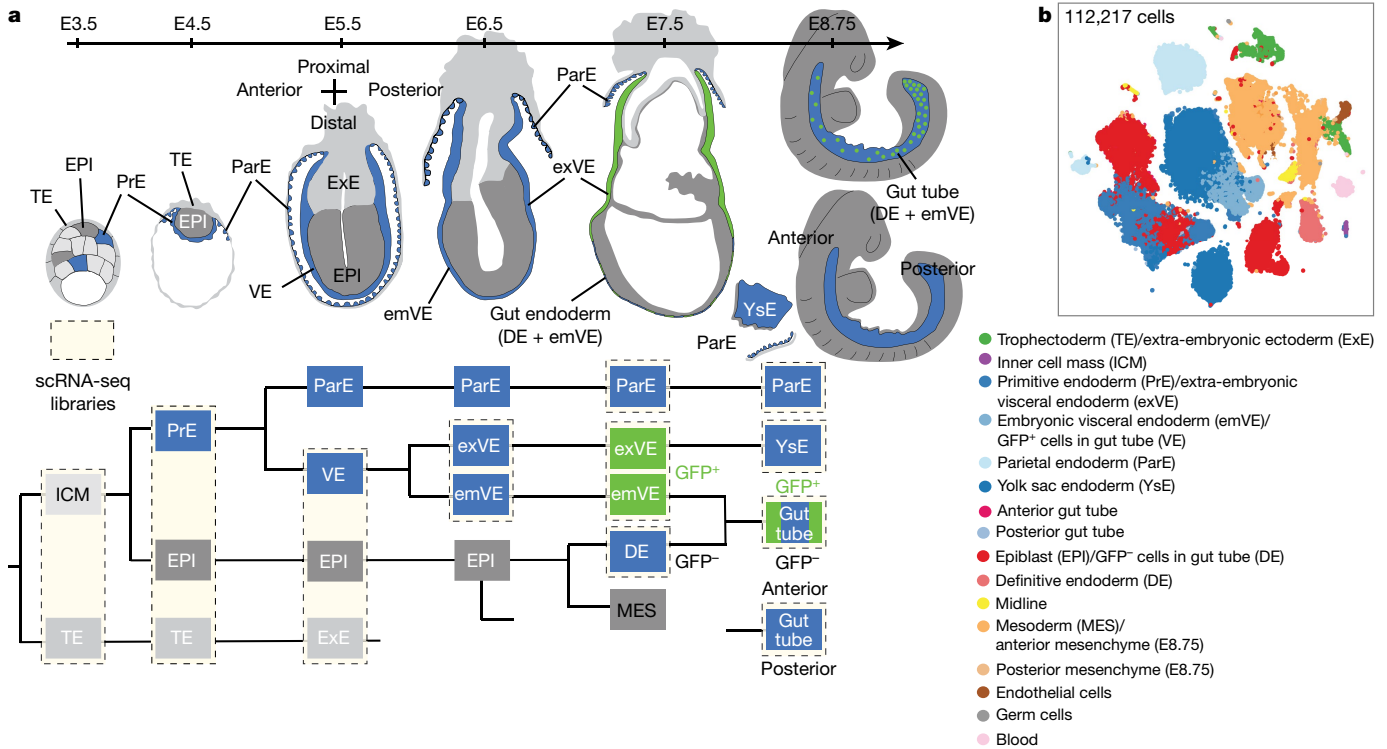


Fig. 1 | Single-cell map of the mouse endoderm, from blastocyst to midgestation. a, Schematic highlighting embryonic stages sampled, lineage relationships and single-cell libraries collected across sequential

stages. **b**, t-SNE plot of all samples; each dot represents a single cell that is colour-coded by cell type. DE, definitive endoderm; TE, trophoctoderm; ParE, parietal endoderm; VE, visceral endoderm; YsE, yolk sac endoderm.

the differentiation potential of a cell as the entropy of its fate probabilities, which represents a measure of the plasticity associated with each state. Regions along pseudo-time at which the differentiation potential drops represent points at which lineage specification and commitment occur.

Emergence of primitive endoderm

The mammalian blastocyst comprises three lineages³: the trophoctoderm, which gives rise to the fetal portion of the placenta; the EPI, which is the progenitor for most somatic tissues, germ cells and extra-embryonic mesoderm; and the PrE, which gives rise to the endodermal component of visceral and parietal yolk sacs, and gut endoderm⁷. Following application of Harmony to E3.5 and E4.5 datasets, force-directed layout illustrates the relationship between blastocyst lineages (Extended Data Fig. 4a, b). On the basis of the average diffusion distance^{10,11} from the bipotent inner cell mass (ICM), trophoctoderm cells were substantially further away from ICM than either EPI or PrE (Extended Data Fig. 4b, Methods). This suggests that the decision between trophoctoderm and ICM is complete before ICM cells make a choice between PrE and EPI, and that EPI cells are phenotypically closer to ICM cells than they are to PrE cells (Extended Data Fig. 4c).

To pinpoint the time point at which fate decisions occur, and to characterize gene expression dynamics during commitment, we excluded trophoctoderm cells and applied Palantir^{10,11} using a *Nanog*^{high} cell (uncommitted ICM) as the start cell (Extended Data Fig. 4d, e). The changes in differentiation potential and branch probabilities suggest that the ICM lineage divergence into EPI and PrE occurs at E3.5, consistent with previous analyses that used limited markers^{20,21} (Extended Data Fig. 4e, f). Two ICM clusters were identified: one cluster represented uncommitted cells that had an equal propensity for PrE and EPI fate (purple in Extended Data Fig. 4d, g, h), and another cluster—although uncommitted—had started to specify towards PrE or EPI (green in Extended Data Fig. 4d, g, h, Supplementary Table 1). We also identified two PrE clusters at E3.5, which we propose represent nascent (light blue in Extended Data Fig. 4d) and more advanced (dark blue in Extended Data Fig. 4d) populations during lineage maturation²².

At E4.5, we observed distinct EPI and PrE populations. Two clusters were identified within the PrE at E4.5 (clusters shown in light and dark blue in Extended Data Fig. 4d, i) that probably represent emergent visceral and parietal endoderm (orange and black arrowheads, respectively, in Extended Data Fig. 4d, i, Supplementary Table 2).

ICM cell-fate specification is driven, in part, by the lineage-specific transcription factors *GATA6* and *NANOG*, which are co-expressed in uncommitted ICM and exclusively expressed in—and required for—PrE and EPI, respectively^{23,24}. Although *FGF4* signalling is active across the ICM population^{25–27}, it is critical for PrE specification^{28,29}. The dynamics of key transcription factors and components of signalling pathways within the ICM remain unclear. We used Palantir to characterize the expression trends for *Gata6*, *Nanog* and *Fgf4* along pseudo-time, as PrE and EPI cells emerged (Extended Data Fig. 5a).

The ratio between *Gata6* and *Nanog* tracked closely with EPI specification, and was a strong descriptor of ICM cell-fate specification (Extended Data Fig. 5b). Expression ratio between *Gata6* and *Fgf4* also tracked with EPI specification, but trailed the ratio between *Gata6* and *Nanog* along the inferred pseudo-time ordering. These analyses provide a precise ordering of markers during ICM lineage specification^{3,22,28} (Extended Data Fig. 5c). Additional genes correlated along pseudo-time with *Fgf4* in EPI (for example, *Tcf7l1*) (Extended Data Fig. 5b, d–f, green arrowheads), and *Gata6* and *Gata4* in PrE (Extended Data Fig. 5b, Supplementary Fig. 3).

Mouse mutants in both *Fgfr1* and *Fgfr2* phenocopy embryos that lack *Fgf4*, and display defects in PrE specification and exit from naive pluripotency in the EPI^{26,27,29}. Palantir analysis suggested that *Fgfr1* was expressed in uncommitted ICM cells (Extended Data Fig. 5a, c, d) and downregulated upon PrE specification, at the time of transient *Fgfr2* activation. This tandem receptor expression suggests sequential FGFR1 and FGFR2 activity during PrE specification (Extended Data Fig. 5c). By E4.5, at which point *Fgfr1* and *Fgfr2* are no longer expressed, a second phase of FGF signalling in PrE could be mediated by FGF5 and/or FGF8 signalling through FGFR4 in emergent visceral endoderm, and FGF3 signalling via FGFR3 parietal endoderm (Extended Data Fig. 5d, orange and black arrowheads). While in EPI, FGF4 and FGFR1 may

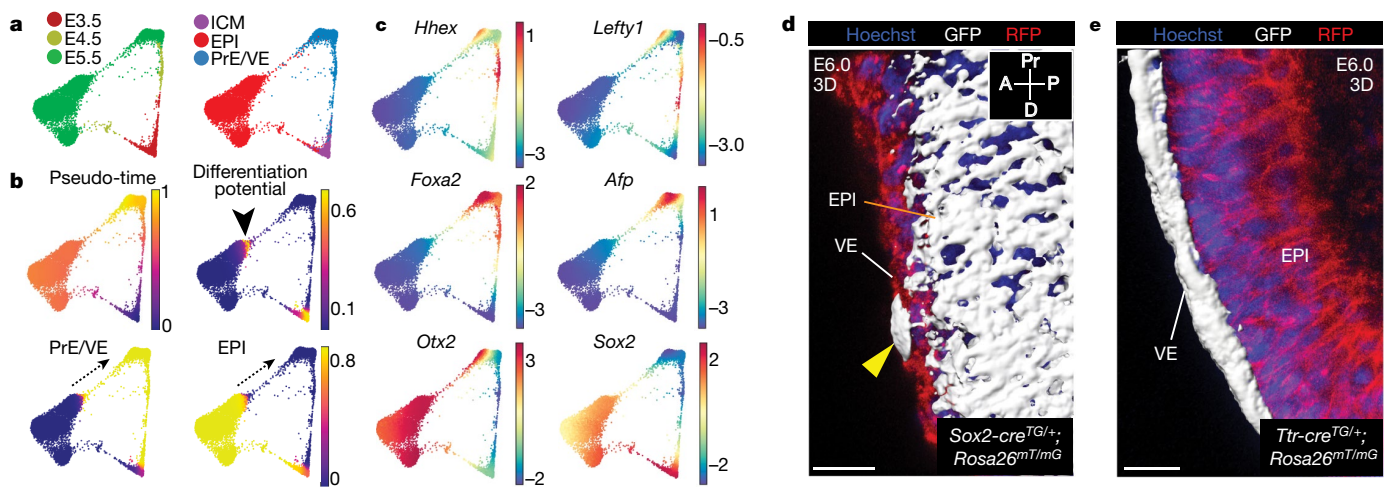


Fig. 2 | Differentiation of EPI into endoderm before gastrulation. Results from Harmony applied to all replicates of E3.5–E5.5 time points. **a**, Force-directed layouts that depict the relationship between EPI, and PrE and visceral endoderm lineages. Cells coloured by time-point (left) and cell-type labels (right). **b**, Palantir pseudo-time, differentiation potential and branch probabilities of EPI, and PrE and visceral endoderm cell lineages. Black arrowhead and dotted arrows denote EPI cells with a high differentiation potential, which represents a trans-differentiation to endoderm. **c**, Gene expression of AVE (*Hhex* and *Lefty1*), visceral

endoderm (*Foxa2* and *Afp*), visceral endoderm and EPI (*Otx2* and *Sox2*) markers. Cells coloured by gene expression post-imputation with the MAGIC algorithm⁵⁰. **d**, **e**, Three-dimensional surface renderings of mGFP-expressing cells in *Sox2-cre*^{TG/+};*Rosa26*^{mT/mG} (**d**) and *Ttr-cre*^{TG/+};*Rosa26*^{mT/mG} embryos (**e**) at E6.0. Nuclei stained with Hoechst and membranes labelled with red fluorescent protein (RFP). Yellow arrowhead indicates a GFP-positive EPI cell that has intercalated into the visceral endoderm. Results validated in more than three independent experiments. Scale bars, 10 μ m. A, anterior; D, distal; P, posterior; Pr, proximal.

be driving pluripotent state transitions (Extended Data Fig. 5a, c, d, green arrowheads).

Differentiation of EPI to endoderm

Although EPI and PrE were distinct at E4.5, by E5.5 we observed a continuum of cells that exhibited a gradual increase in expression of endodermal marker genes that bridged between the EPI and visceral endoderm (Fig. 2a, Extended Data Fig. 6a, black arrowhead, c, d). By contrast, no connection was observed between EPI or visceral endoderm and extra-embryonic ectoderm (ExE), which is a descendant of the trophoblast. On the basis of average pairwise distances, ExE cells were phenotypically more distinct from EPI and visceral endoderm at E5.5, as compared to PrE (Extended Data Fig. 6a–d). Investigation of gene trends within the EPI and visceral endoderm (Supplementary Fig. 4) identified genes that correlate with endoderm factors (such as *Foxa2*, *Gata4*, *Gata6*, *Sox7* and *Sox17*) as well as pluripotency-associated factors (such as *Nanog*, *Pou3f1* and *Klf4*) (Supplementary Table 3).

To define the crossover between EPI and visceral endoderm, we applied Palantir to Harmony-augmented data from cells at E3.5–E5.5, after excluding trophoblast and ExE cells (Fig. 2b) and using the same *Nanog*^{high} start cell as in Extended Data Fig. 4. Palantir ordered cells along their developmental trajectories with high differentiation potential at E3.5, which corresponds to the lineage divergence into EPI and PrE (Fig. 2b). An increase in differentiation potential was also observed in a subset of EPI cells that bridged to the visceral endoderm at E5.5 (Fig. 2b, black arrowhead). After the region of high differentiation potential, we observed a sharp increase in the visceral endoderm branch probabilities, which indicates that the bridging cells had high propensity towards the visceral endoderm fate (Fig. 2b, arrows with dotted lines). This suggests that the continuity between EPI and visceral endoderm results from a subset of EPI cells that acquire an endoderm identity. The expression of markers (such as *Lefty1*, *Cer1* and *Hhex*^{30–32}) of embryonic visceral endoderm (emVE) and anterior visceral endoderm (AVE)—a specialized cellular cohort of emVE that exhibits an intra-epithelial, distal-to-anterior migratory behaviour between E5.5 and E6.0—suggested these EPI descendants resembled emVE and AVE cells³³ (Fig. 2c, Extended Data Fig. 6e).

To validate the crossover between EPI and visceral endoderm, we used two in vivo lineage-tracing approaches. First, we crossed the EPI-specific *Sox2-cre*³⁴ and visceral endoderm-specific *Ttr-cre*³⁵ mouse

lines to the *Rosa26*^{mtdTomato/mGFP(mTmG)} (ref. 36) reporter line (Fig. 2d, e, Extended Data Fig. 6f, g). Imaging of *Sox2-cre*^{TG/+};*Rosa26*^{mTmG/+} embryos at E5.5–E6.0 revealed that the majority of GFP-positive cells were within the EPI, and that single GFP-positive cells were also present within the emVE (yellow arrowheads in Extended Data Fig. 6f, g; 1–5 cells, in 10 out of 20 embryos) but not within the extra-embryonic visceral endoderm (exVE). Transmigrating cells were GATA6-positive, which indicates they had acquired an endoderm identity. At E5.5, GFP-positive, GATA6-positive cells were observed in distal locations, whereas by E6.0 these cells predominantly resided more anteriorly (Extended Data Fig. 6f). By contrast, all GFP-positive cells

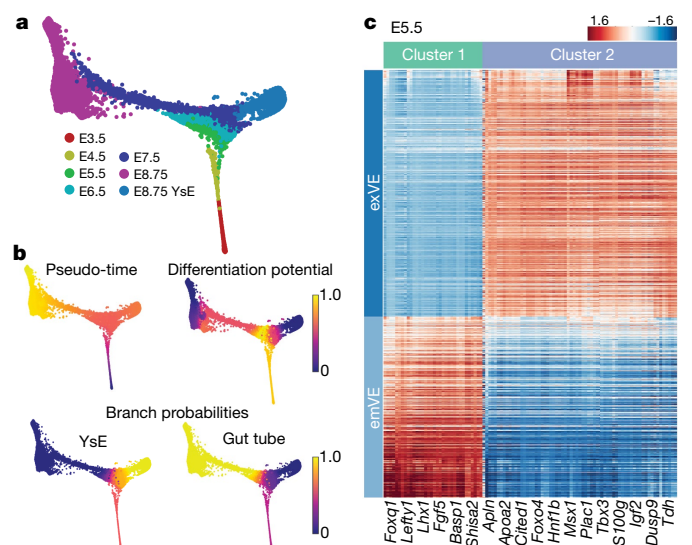


Fig. 3 | Spatial pattern emerges within visceral endoderm at the onset of post-implantation development at E5.5. Results from Harmony applied to replicates of E3.5–E8.75 time points (excluding parietal endoderm). **a**, Force-directed layout of endoderm cells from blastocyst to midgestation. **b**, Palantir pseudo-time, differentiation potential and branch probabilities of endoderm cells using a *Nanog*^{high} start cell. **c**, Heat map of genes expressed (Extended Data Fig. 9h) in exVE or emVE at E5.5. Cells are sorted within each compartment by pseudo-time ordering.

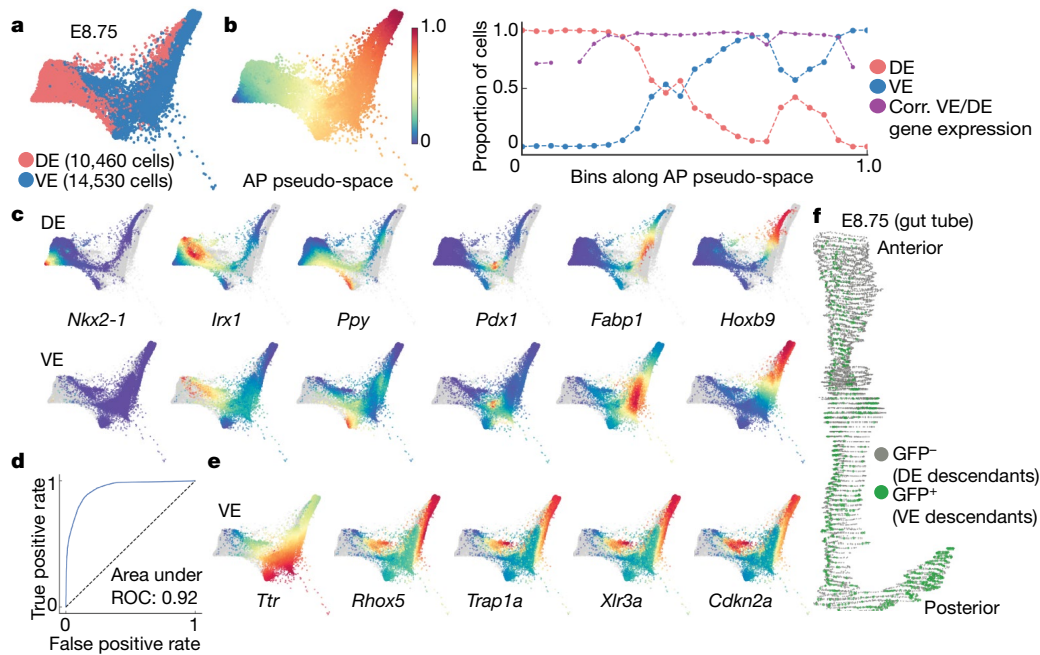


Fig. 4 | Anterior–posterior pseudo-spatial axis of cells that reside within the gut tube at E8.75. a–c, e, Force-directed layout of E8.75 visceral and definitive endoderm cells that combines anterior and posterior cells with AFP–GFP-positive and AFP–GFP-negative cells, using mNNCorrect¹⁸. **a,** Cells coloured based on measured or inferred anterior–posterior position. **b,** Inferred anterior–posterior (AP) pseudo-space (left) and proportion of visceral and definitive endoderm cells in bins along the anterior–posterior pseudo-space (right). Purple dots represent correlation

were restricted to the visceral endoderm in *Ttr-cre*^{TG/+}; *Rosa26*^{mT/mG} embryos, and no cells were detected in the EPI (0 out of 27 embryos, Fig. 2e, Extended Data Fig. 6f). We also generated chimaeras between tetraploid wild-type embryos and *CAG-H2B-tdTomato* embryonic stem cells, and observed that tdTomato-positive cells were distributed throughout the EPI and sparsely within the emVE (1–5 cells, in 9 out of 19 embryos, Extended Data Fig. 6g).

Embryonic and extra-embryonic visceral endoderm

The early post-implantation (E5.5, Fig. 1) mouse embryo is radially symmetrical, and the visceral endoderm appears morphologically uniform around its proximal–distal axis^{4,33}. Symmetry is broken, and the anterior–posterior axis established, through the migration of AVE cells³⁷. Proximal–distal spatial patterning across the visceral endoderm has previously been described as preceding, and coincident with, the onset of gastrulation at E6.5. There is a clear distinction between the morphology and function of the proximally located exVE, a cuboidal epithelium that overlies the ExE and that gives rise to yolk sac endoderm, and distal emVE, a squamous epithelium that overlies the EPI and that contributes to the gut tube⁷. To determine the onset of transcriptomic determinants of spatial patterning within the visceral endoderm, we sought to establish at which point cells that are specified as yolk sac endoderm and gut tube are identified. We used Harmony to integrate cells of the visceral endoderm lineage at E3.5–E8.75 (Fig. 3a), and applied Palantir using a *Nanog*^{high} ICM start cell (Fig. 3b, Extended Data Fig. 7a, b).

A clear distinction was evident between cells that specify towards yolk sac endoderm versus gut tube at E6.5 and E7.5 (Supplementary Fig. 5a). Consistent with reported spatial patterning at these stages, cells that specify towards yolk sac endoderm and gut tube were identified as exVE and emVE cells, respectively, on the basis of marker expression in scRNA-seq data, and correlations with bulk RNA-seq expression of sorted exVE and emVE tissues at E7.5 (Extended Data Fig. 7c–e). Visceral endoderm cells at E5.5 were distributed across pseudo-time (Extended Data Fig. 7a): a subset of these cells did not

exhibit any change in differentiation potential, which indicates a more uncommitted state (Extended Data Fig. 7e, f), whereas a majority of these cells exhibited an altered differentiation potential, which indicates their propensity towards either a yolk sac endoderm or gut tube fate (Extended Data Fig. 7e, f). These data reveal that, at the transcriptional level, spatial patterning exists at E5.5 and precedes morphological changes within the visceral endoderm.

Differential expression between bulk RNA-sequenced exVE and emVE populations at E7.5 suggested that emVE represents a specialized variant of exVE (Extended Data Fig. 7d), which perhaps modulates a transcriptional program in response to stimuli, such as BMP or NODAL^{38–40}. To explore this further, we identified two covarying gene sets (Supplementary Fig. 5b) that exhibit contrasting expression patterns in putative exVE and emVE cells at E5.5 (Fig. 3c) and distinguish bona fide exVE and emVE cells at E6.5 and E7.5 (Extended Data Fig. 7g). EmVE-specific genes (cluster 1, Fig. 3c) included *Lhx1* and *Lefty1*, and the AVE-specific genes *Cer1* and *Hhex*^{30,32,37}. ExVE-specific genes (cluster 2, Fig. 3c) included *Apln* and *Msx1* (Supplementary Table 4). Whole-mount in situ hybridization (ISH) for *Apln*, *Lhx1*, *Lefty1* and *Msx1* at E6.25 validated regionalized expression (Extended Data Fig. 7h). These data demonstrate that the visceral endoderm is patterned at the onset of post-implantation development, and that the emVE cells—including the AVE subpopulation—are derivative of exVE.

Anterior–posterior patterning of gut endoderm

We combined data from anterior and posterior gut tube compartments at E8.75, with GFP-positive and GFP-negative populations (Fig. 4a), using a manifold classifier (Supplementary Note 3) to infer the GFP status of anterior and posterior cells, and the anterior–posterior position of GFP-positive and GFP-negative cells (Extended Data Fig. 8a, b). The strongest signal in the data, as determined by the first diffusion component, was cell ordering along the anterior–posterior axis (Extended Data Fig. 8c). To corroborate that anterior–posterior ordering reflects spatial distribution along the gut tube, we confirmed

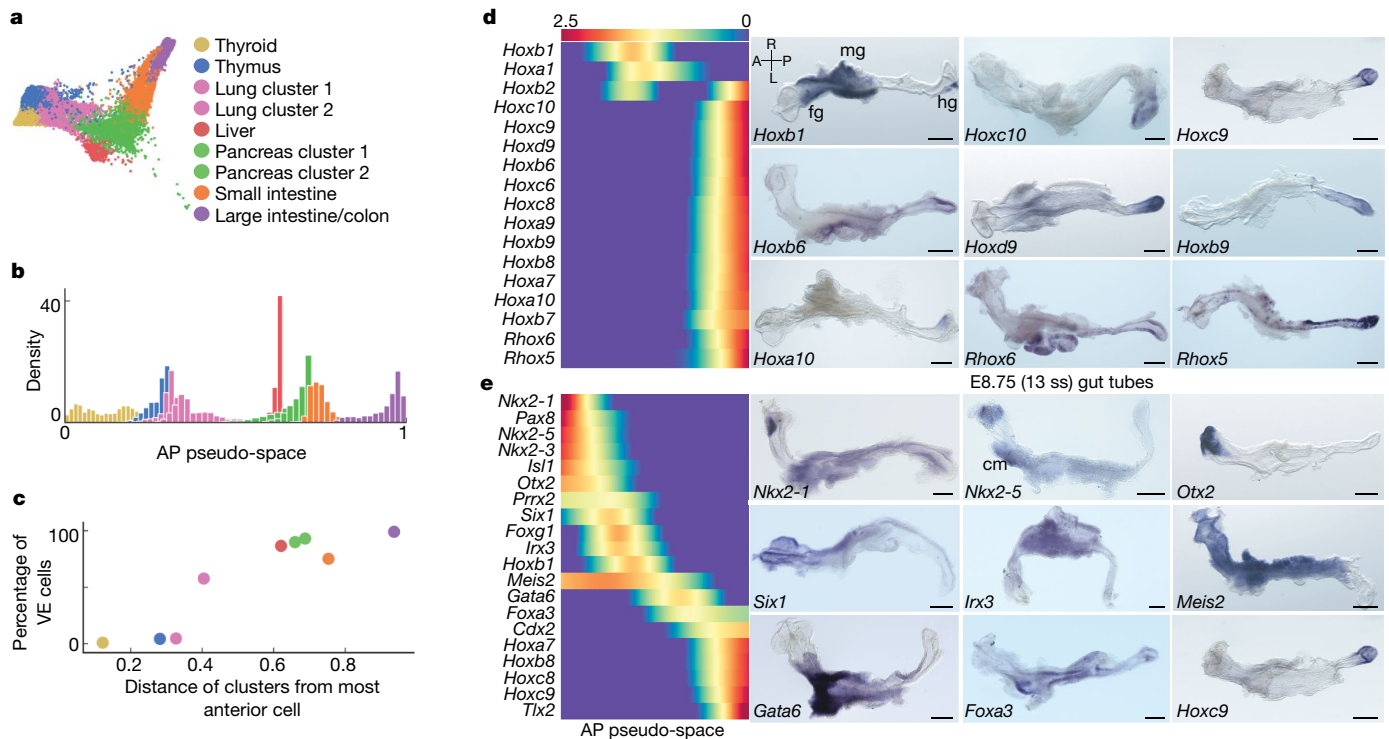


Fig. 5 | Spatial patterning and organ identities within the gut tube of the mouse embryo at E8.75. **a**, Force-directed graph of cells at E8.75, coloured by Phenograph clusters, annotated with the putative endodermal organ that is associated with each cluster. **b**, Density of cells, per Phenograph cluster, along the anterior–posterior pseudo-space. **c**, Percentage of visceral endoderm cells per cluster, ordered by average distance from anterior tip of anterior–posterior pseudo-space. **d**, Heat map of Hox gene expression along the anterior–posterior pseudo-space (left). Validation of Hox gene expression by ISH on gut tubes at E8.75 ($n > 3$ for each gene) (right). Scale bars, 100 μm (*Hoxc10*), 200 μm (all others). ss, somite

stage. **e**, Heat map of transcription factor expression most predictive of anterior–posterior pseudo-space in a regression model. Columns represent cells ordered by pseudo-space; each row represents the expression of a particular transcription factor. Transcription factors are ordered by their expression along the anterior–posterior pseudo-space. Validation of predictive anterior–posterior expression by ISH on gut tubes at E8.75 ($n > 3$ for each gene) (right). Scale bars, 100 μm (*Nkx2-1*, *Irx3*), 200 μm (all others). cm, cardiac mesoderm; fg, foregut; hg, hindgut; L, left; mg, midgut; R, right.

consistency between gene expression trends from scRNA-seq data and bulk RNA-seq data of micro-dissected gut tube quadrants, such that *Nkx2-1* (a gene expressed in the anterior gut tube) and *Hoxb9* (a gene expressed in the posterior gut tube) exhibited consistent expression patterns (Extended Data Fig. 8c). To determine a more robust ordering, we inferred pseudo-spatial ordering of cells along the gut tube by computing multiscale distances from the anterior-most cell after projecting cells onto multiple diffusion components (Fig. 4b, Extended Data Fig. 8d, Supplementary Note 4). The inferred pseudo-space was robust to different parameters, and reproducible across replicates (Extended Data Fig. 8e–g).

Gut endoderm comprises EPI-derived definitive and visceral endoderm descendants^{6,7}. Consistent with previous findings, we observed extensive intermixing of these descendants along the anterior–posterior pseudo-space axis, with enrichment of definitive endoderm descendants in the anterior and visceral endoderm descendants in the posterior region (Fig. 4b). To determine whether visceral endoderm descendants attained transcriptional equivalence with definitive endoderm descendants, we compared the expression of markers of the emergent endodermal organs within both populations: *Nkx2-1* (thyroid and thymus)⁴¹, *Irx1* (lung)⁴², *Ppy* (liver)⁴³, *Pdx1* (pancreas)⁴⁴, *Fabp1* (small intestine)⁴⁵ and *Hoxb9* (posterior gut tube). All genes were expressed at substantial levels in both visceral and definitive endoderm cells, and at comparable anterior–posterior positions—except for *Nkx2-1*, which is expressed in the anterior-most cells of the gut tube and is therefore exclusive to the definitive endoderm (Fig. 4c). Furthermore, we noted a strong correlation in global gene expression patterns between visceral and definitive endoderm cells in bins along the anterior–posterior pseudo-space (Fig. 4b, purple), which suggests that they were patterned similarly to one another and acquired regionalized organ-specific identities.

A memory of extra-embryonic lineage history

Despite the global similarity in transcriptomes, visceral and definitive endoderm descendants might retain a memory of their lineage history. To overcome the confounding effects that are introduced by the spatial distribution of visceral and definitive endoderm descendants along the gut tube, we trained a sparse logistic regression model to classify visceral and definitive endoderm cells at E7.5, using all genes as features (Supplementary Note 5). This classifier achieved near-perfect accuracy on a test set of E7.5 cells (area under receiving operating characteristic curve (auROC) 0.96) (Extended Data Fig. 8h). We applied this classifier (trained on data from E7.5 cells) to predict the origin of cells within the gut tube at E8.75, and achieved a similarly high accuracy (Fig. 4d, auROC 0.92). Thus, despite the extensive morphological and transcriptional changes that take place between E7.5 and E8.75, visceral endoderm lineage history is maintained through the expression of a core set of genes, including *Rhox5*, *Trap1a*, *Xlr3a*, *Cdkn2a* and *Ttr* (Fig. 4e, Extended Data Fig. 8i, j, Supplementary Table 5).

Emergence of organ identities

To determine whether the emergence of spatial patterning along the anterior–posterior axis could be observed earlier in development, we applied Palantir separately to definitive and visceral endoderm cells. Our results revealed the presence of a small fraction of cells that acquires anterior–posterior identity in both definitive and visceral endoderm compartments at E7.5 (Extended Data Fig. 8k). Notably, definitive and visceral endoderm cells were predominantly primed towards anterior and posterior localization, respectively (black arrowheads in Extended Data Fig. 8k). We next compared the distribution of visceral endoderm cells along the anterior–posterior axis of embryos, with visceral endoderm proportions inferred from the scRNA-seq data

along the pseudo-space axis. To quantify the distribution of visceral and definitive endoderm cells within embryo gut tubes at E8.75 (13 somite stage) at cellular resolution, we analysed serial transverse sections of three *Afp-GFP^{TG/+}* embryos (Fig. 4f, Extended Data Fig. 8l). Visceral endoderm descendant proportions in binned locations along the anterior–posterior axis and anterior–posterior pseudo-space axis were highly correlated (Extended Data Fig. 8m), which further demonstrates the accuracy of inferred anterior–posterior pseudo-space.

To investigate whether the gut tube at E8.75 already contains information that relates to later organ establishment, we clustered all cells, annotated clusters on the basis of differential expression of primordial organ markers and determined an ordering of clusters along the anterior–posterior pseudo-space. The resulting ordering of clusters matched the sequence of organ identities along the anterior–posterior axis of the gut tube (Fig. 5a). We observed a high degree of variability in the density of cells along the pseudo-spatial axis (Fig. 5b, c, Extended Data Fig. 9a, b), with low-density regions between clusters. Cluster-specific expression was validated using ISH, which confirmed the accuracy of the inferred anterior–posterior pseudo-space as well as the emergence of endodermal organ identities at E8.75 (Extended Data Fig. 9a).

Hox gene expression in the developing central nervous system is considered a canonical descriptor of anterior–posterior axis position⁴⁶. Although not all Hox genes were expressed within the gut tube at E8.75, the majority of Hox genes that were expressed were posteriorly localized (Fig. 5d, Extended Data Fig. 10)—including several genes that displayed robust, more anterior expression within the mesoderm and/or neuroectoderm (*Hoxb1* and *Hoxd4*) (Extended Data Fig. 10). These data suggest that anterior–posterior patterning of the gut tube, and the stereotypical emergence of organ identities, precede or are independent of a Hox code.

To generate a signalling map of the gut tube (Extended Data Fig. 11), we analysed the expression of context-independent targets for the activity of key signalling pathways. Our data validate FGF and WNT, and reveal NOTCH signalling, at the posterior of the gut tube (small and large intestine clusters, Fig. 5). BMP, HH, JAK/STAT and HIPPO pathway activation encompassed multiple domains; NODAL signalling was not active at this stage; and both positive and negative read-outs of retinoic acid signalling were posteriorly localized.

To examine the contribution of cell-autonomous cues to the anterior–posterior pattern within the gut tube, we trained a sparse regression model to predict anterior–posterior pseudo-space using the expression of all transcription factors as features (Supplementary Note 6). Transcription factor expression was exceptionally accurate in predicting anterior–posterior pseudo-space order (Extended Data Fig. 9c, correlation 0.97), which indicates that transcriptional regulation—presumably in response to signals from neighbouring mesenchyme¹—has a key role in gut tube patterning. This model identified a core group of 20 transcription factors that predict the anterior–posterior pseudo-space ordering (Fig. 5e, Extended Data Fig. 9d–f). Expression domains for the core factors—from *Nkx2-1* at the anterior to *Hoxb9* and *Hoxc9* at the posterior—were validated by ISH (Fig. 5e).

Discussion

We have delineated the transcriptional landscape of mouse endoderm from pre-implantation to midgestation. Our data pinpoint the order and timing of key events that start with the emergence of the primitive endoderm population within the blastocyst. Our data also define previously unappreciated sub-states within well-studied cell populations, and uncovered detailed gene expression trends.

The analysis reveals that, throughout embryogenesis, cells acquire a transcriptional identity that reflects their future fate and spatial positioning before overt spatial organization. For example, there is transcriptional priming of the spatial patterning of cells along the anterior–posterior axis at E7.5. Although cells develop a marked propensity towards specific cell fates earlier than previously appreciated, they nevertheless retain a notable degree of plasticity. Application of

Palantir to our data suggested plasticity within the EPI lineage that was validated through lineage-tracing experiments, in which the EPI differentiates into endoderm before the onset of gastrulation. One might speculate whether this EPI-to-endoderm differentiation reflects a removal of ‘less-fit’ cells from the pluripotent compartment, or an active recruitment of cells to the visceral endoderm. In the context of cell competition-based models for the EPI, cell engulfment or apical cell extrusion have previously been proposed as mechanisms for cell removal⁴⁷. In considering an active recruitment of EPI cells to visceral endoderm, it has previously been suggested that breaks in the basement membrane at the interface between the EPI and visceral endoderm might allow cells to escape the EPI layer, and populate the nascent AVE⁴⁸.

At gastrulation (E7.0–E7.5), EPI-derived definitive endoderm cells intercalate into the emVE epithelium to form the gut endoderm^{5,7}. Although we show that visceral and definitive endoderm descendants retain a signature of their lineage history, the data suggest they largely acquire transcriptomic equivalence. By the time the gut endoderm has internalized (forming the gut tube (E8.75)), clusters of cells that express markers of organ identity were identified, and correlated in anterior–posterior pseudo-space with the stereotypical order of endodermal organs. These were determined largely by spatial localization instead of lineage history.

Cell fate is determined through a combination of cell-intrinsic propensities to specific fates, and extrinsic cues (for example, signalling) from the environment; detailed knowledge of these inputs should yield improved protocols for differentiation into distinct endodermal derivatives⁴⁹. A future challenge will be to dissect the dynamic interplay between different inputs in determining the coordination of fate decisions that underlies the emergence of distinct organ identities of defined size at stereotypical locations along the gut tube, and to investigate the persistence and function of visceral endoderm descendants within endodermal organs.

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/s41586-019-1127-1>.

Received: 14 November 2018; Accepted: 29 March 2019;

Published online 8 April 2019.

- Zorn, A. M. & Wells, J. M. Vertebrate endoderm development and organ formation. *Annu. Rev. Cell Dev. Biol.* **25**, 221–251 (2009).
- Tremblay, K. D. Formation of the murine endoderm: lessons from the mouse, frog, fish, and chick. *Prog. Mol. Biol. Transl. Sci.* **96**, 1–34 (2010).
- Chazaud, C. & Yamanaka, Y. Lineage specification in the mouse preimplantation embryo. *Development* **143**, 1063–1074 (2016).
- Nowotschin, S. & Hadjantonakis, A. K. Cellular dynamics in the early mouse embryo: from axis formation to gastrulation. *Curr. Opin. Genet. Dev.* **20**, 420–427 (2010).
- Viotti, M., Nowotschin, S. & Hadjantonakis, A. K. SOX17 links gut endoderm morphogenesis and germ layer segregation. *Nat. Cell Biol.* **16**, 1146–1156 (2014).
- Viotti, M., Nowotschin, S. & Hadjantonakis, A. K. *Afp:mCherry*, a red fluorescent transgenic reporter of the mouse visceral endoderm. *Genesis* **49**, 124–133 (2011).
- 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).
- Sherwood, R. I., Chen, T. Y. & Melton, D. A. Transcriptional dynamics of endodermal organ formation. *Dev. Dyn.* **238**, 29–42 (2009).
- Hou, J. et al. A systematic screen for genes expressed in definitive endoderm by Serial Analysis of Gene Expression (SAGE). *BMC Dev. Biol.* **7**, 92 (2007).
- Setty, M. et al. Palantir characterizes cell fate continuities in human hematopoiesis. Preprint at <https://www.biorxiv.org/content/early/2018/08/05/385328> (2018).
- Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).
- Kwon, G. S. et al. *Tg(Afp-GFP)* expression marks primitive and definitive endoderm lineages during mouse development. *Dev. Dyn.* **235**, 2549–2558 (2006).
- Azizi, E. et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293–1308 (2018).

14. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
15. Amir, E.-a. D. et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–552 (2013).
16. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).
17. 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).
18. Haghverdi, L., Büttner, M., Wolf, F. A., Büttner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
19. Farrell, J. A. et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131 (2018).
20. Plusa, B., Piliszek, A., Frankenberg, S., Artus, J. & Hadjantonakis, A. K. Distinct sequential cell behaviours direct primitive endoderm formation in the mouse blastocyst. *Development* **135**, 3081–3091 (2008).
21. Chazaud, C., Yamanaka, Y., Pawson, T. & Rossant, J. Early lineage segregation between epiblast and primitive endoderm in mouse blastocysts through the Grb2-MAPK pathway. *Dev. Cell* **10**, 615–624 (2006).
22. Artus, J., Piliszek, A. & Hadjantonakis, A. K. The primitive endoderm lineage of the mouse blastocyst: sequential transcription factor activation and regulation of differentiation by Sox17. *Dev. Biol.* **350**, 393–404 (2011).
23. Silva, J. et al. Nanog is the gateway to the pluripotent ground state. *Cell* **138**, 722–737 (2009).
24. Schrode, N., Saiz, N., Di Talia, S. & Hadjantonakis, A. K. GATA6 levels modulate primitive endoderm cell fate choice and timing in the mouse blastocyst. *Dev. Cell* **29**, 454–467 (2014).
25. Morgani, S. M. et al. A *Sprouty4* reporter to monitor FGF/ERK signaling activity in ESCs and mice. *Dev. Biol.* **441**, 104–126 (2018).
26. Molotkov, A., Mazot, P., Brewer, J. R., Cinali, R. M. & Soriano, P. Distinct requirements for FGFR1 and FGFR2 in primitive endoderm development and exit from pluripotency. *Dev. Cell* **41**, 511–526 (2017).
27. Kang, M., Garg, V. & Hadjantonakis, A. K. Lineage establishment and progression within the inner cell mass of the mouse blastocyst requires FGFR1 and FGFR2. *Dev. Cell* **41**, 496–510 (2017).
28. Ohnishi, Y. et al. Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nat. Cell Biol.* **16**, 27–37 (2014).
29. Kang, M., Piliszek, A., Artus, J. & Hadjantonakis, A. K. FGF4 is required for lineage restriction and salt-and-pepper distribution of primitive endoderm factors but not their initial expression in the mouse. *Development* **140**, 267–279 (2013).
30. Thomas, P. Q., Brown, A. & Beddington, R. S. *Hex*: a homeobox gene revealing peri-implantation asymmetry in the mouse embryo and an early transient marker of endothelial cell precursors. *Development* **125**, 85–94 (1998).
31. Meno, C. et al. Mouse Lefty2 and zebrafish Antivin are feedback inhibitors of nodal signaling during vertebrate gastrulation. *Mol. Cell* **4**, 287–298 (1999).
32. Belo, J. A. et al. Cerberus-like is a secreted factor with neutralizing activity expressed in the anterior primitive endoderm of the mouse gastrula. *Mech. Dev.* **68**, 45–57 (1997).
33. Arnold, S. J. & Robertson, E. J. Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nat. Rev. Mol. Cell Biol.* **10**, 91–103 (2009).
34. Hayashi, S., Lewis, P., Pevny, L. & McMahon, A. P. Efficient gene modulation in mouse epiblast using a *Sox2Cre* transgenic mouse strain. *Mech. Dev.* **119**, S97–S101 (2002).
35. Kwon, G. S. & Hadjantonakis, A. K. Transthyretin mouse transgenes direct RFP expression or Cre-mediated recombination throughout the visceral endoderm. *Genesis* **47**, 447–455 (2009).
36. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).
37. Takaoka, K., Yamamoto, M. & Hamada, H. Origin and role of distal visceral endoderm, a group of cells that determines anterior–posterior polarity of the mouse embryo. *Nat. Cell Biol.* **13**, 743–752 (2011).
38. Paca, A. et al. BMP signaling induces visceral endoderm differentiation of XEN cells and parietal endoderm. *Dev. Biol.* **361**, 90–102 (2012).
39. Kruithof-de Julio, M. et al. Regulation of extra-embryonic endoderm stem cell differentiation by Nodal and Cripto signaling. *Development* **138**, 3885–3895 (2011).
40. Artus, J. et al. BMP4 signaling directs primitive endoderm-derived XEN cells to an extraembryonic visceral endoderm identity. *Dev. Biol.* **361**, 245–262 (2012).
41. Serra, M. et al. Pluripotent stem cell differentiation reveals distinct developmental pathways regulating lung- versus thyroid-lineage specification. *Development* **144**, 3879–3893 (2017).
42. Becker, M. B., Zülch, A., Bosse, A. & Gruss, P. *Irx1* and *Irx2* expression in early lung development. *Mech. Dev.* **106**, 155–158 (2001).
43. Yang, Y., Akinci, E., Dutton, J. R., Banga, A. & Slack, J. M. Stage specific reprogramming of mouse embryo liver cells to a beta cell-like phenotype. *Mech. Dev.* **130**, 602–612 (2013).
44. Offield, M. F. et al. PDX-1 is required for pancreatic outgrowth and differentiation of the rostral duodenum. *Development* **122**, 983–995 (1996).
45. Tsai, Y. H. et al. *In vitro* patterning of pluripotent stem cell-derived intestine recapitulates *in vivo* human development. *Development* **144**, 1045–1055 (2017).
46. Deschamps, J. & van Nes, J. Developmental regulation of the Hox genes during axial morphogenesis in the mouse. *Development* **132**, 2931–2942 (2005).
47. Di Gregorio, A., Bowling, S. & Rodriguez, T. A. Cell competition and its role in the regulation of cell fitness from development to cancer. *Dev. Cell* **38**, 621–634 (2016).
48. Hiramatsu, R. et al. External mechanical cues trigger the establishment of the anterior-posterior axis in early mouse embryos. *Dev. Cell* **27**, 131–144 (2013).
49. McCauley, H. A. & Wells, J. M. Pluripotent stem cell-derived organoids: using principles of developmental biology to grow human tissues in a dish. *Development* **144**, 958–962 (2017).
50. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).

Acknowledgements We thank K. Anderson, A. Joyner, A. Martinez-Arias and L. Mazutis for discussions; L. Beccari, D. Duboule, L. Sussel, M. Torres, D. Wellik and M. Wilkinson for plasmids; B. Merrill for antibodies; and J. Brickman for embryonic stem cells. This work was supported by grants from the NIH (R01-DK084391 and R01-HD094868 to A.-K.H.; DP1-HD084071 and R01-CA164729 to D.P.; P30-CA008748 to C. Thompson), MSKCC Society for Special Projects and Functional Genomics Initiative (to A.-K.H. and D.P.) and NSERC (RGPIN-2018-05018 to P.A.H.). C.S.S. is supported by a SYSTEM postdoctoral training award from the Center for Stem Cell Biology MSKCC.

Reviewer information Nature thanks Thorsten Boroviak, Valerie Wilson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.N., M.S., D.M.C., P.A.H., A.-K.H. and D.P. conceived the research. S.N., Y.-Y.K. and V.G. collected cells from embryos with assistance from C.S.S., N.S. and A.-K.H. S.N., Y.-Y.K. and R.G. performed FACS experiments. S.N., V.G. and S.C.B. generated scRNA-seq libraries. S.C.B. and D.M.C. performed RNA-seq. M.S. and D.P. conceived and developed the Harmony and Palantir algorithms. M.S., V.L. and R.S. performed computational analyses of sequence data. V.G. performed blastocyst immunofluorescence. S.N. and Y.-Y.K. performed lineage-tracing, immunofluorescence and ISH on post-implantation embryos and gut tubes. S.N., M.S., A.-K.H. and D.P. analysed and interpreted the data, and wrote the manuscript with input from all authors.

Competing interests S.C.B. and D.M.C. are employees and shareholders at 10x Genomics.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1127-1>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1127-1>.

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

Correspondence and requests for materials should be addressed to A.-K.H. or D.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

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

Ethical compliance. Mice were maintained in accordance with guidelines from Memorial Sloan Kettering Cancer Center (MSKCC) Institutional Animal Care and Use Committee (IACUC) under protocol no. 03-12-017 (principal investigator A.-K.H.).

Mouse husbandry. Mouse strains used: wild-type CD-1 (Charles River), B6D2F1 (Jackson Laboratory), *Afp-GFP^{TG/+}* (ref. 12), *Sox2-cre^{TG/+} (Edi13^{Tg(Sox2-cre)1Amc/J})³⁴*, *Ttr-cre^{TG/+}* (ref. 35) and *Rosa26^{mt/mG} (Gt(Rosa)26Sor^{tm4}(ACTB-tdTomato,-EGFP)^{Luo/J})³⁶*.

Embryo collection. Mice were housed under a 12-h light/dark cycle. Natural mating was set up between males and 4–6-week-old virgin females, with noon of the day of vaginal plug considered to be E0.5. Pre-implantation embryos were flushed from uterine horns at E3.5 and E4.5 with flushing and holding medium (FHM, Millipore), as previously described⁵¹. Zona pellucidae were removed from blastocysts at E3.5 by incubation in acidic Tyrode's solution (Millipore) at 37 °C for 2–3 min. Embryos were subsequently washed through 2–3 drops of FHM and kept in drops of FHM covered with mineral oil (Sigma) on ice, before cell dissociation. Post-implantation embryos were dissected in DMEM/F12, 5% newborn calf serum (Life Technologies) and staged according to Downs and Davies⁵² or by somite number.

Tetraploid embryo chimaeras. Tetraploid embryo chimaeras were generated at the NYU Rodent Genetic Engineering Core Facility. Three-to-four-week-old female B6D2F1 mice (Jackson Laboratories) were super-ovulated with 5IU PMSG and 5IU hCG at 48-h intervals, and then mated individually to B6D2F1 males. Zygotes were collected at E0.5. After overnight culture in KSOM/AA (Millipore) at 37 °C in an atmosphere of 5% CO₂, 2-cell-stage embryos were washed in 0.3 M D-mannitol plus 0.3% BSA (Sigma) and transferred to a Fusion Electrode slide (GSS-250, BLS), and pulses of 30 V for 30 μs were applied. Embryos were monitored for fusion every 30 min. Embryos in which fusion had occurred were cultured for 48 h, until they developed into blastocysts. H2B-tTomato-expressing embryonic stem cells⁵³ were injected into tetraploid blastocysts, and injected blastocysts were cultured to allow for recovery of morphology, before transfer into uteri (up to ten embryos per horn) of E2.5 pseudo-pregnant females (CD-1, Charles River) using standard protocols⁵¹. Chimeric embryos were recovered at E5.5–E6.0 (Extended Data Fig. 6f).

Dissociation of embryos and collection of single cells. 13ss (E8.75)—approximately corresponding to midgestation, the latest stage analysed in this study—is the latest stage for unambiguous assignment of visceral-versus-definitive endoderm origin of gut tube cells using *Afp-GFP^{TG/+}* or *Ttr-cre^{TG/+}* mouse lines. To obtain single cells from 13ss gut tubes, *Afp-GFP^{TG/+}* embryos were dissected, with extra-embryonic membranes and heads removed. Torsos were washed in three drops of DMEM/F12 on ice and incubated in pancreatin/trypsin (2.5% pancreatin/0.5% trypsin in PBS) for 5 min (the exact time was batch-dependent and empirically tested) on ice, and then washed in three drops of DMEM/F12, 10% newborn calf serum on ice. Gut tubes were isolated using tungsten needles (FST cat. no. 10130-10) and washed in ice-cold DMEM/F12. Gut tubes were incubated for 20 min at 37 °C in accutase/0.25% trypsin (1:2) for dissociation into single cells. To obtain single cells from E7.5 definitive endoderm and visceral endoderm, and E6.5 visceral endoderm, embryos were washed in three drops of DMEM/F12 on ice and incubated in pancreatin/trypsin (2.5% pancreatin/0.25% trypsin in PBS) for 3 min (E7.5) and 45 s (E6.5) on ice, and then washed in three drops of DMEM/F12, 10% newborn calf serum on ice. The endoderm layer was teased apart using tungsten needles and washed in cold DMEM/F12, then incubated for 20 min at 37 °C in accutase/0.25% trypsin (1:2). For E5.5 (defined as the stage at which the AVE is distally positioned, observed as a thickening within the emVE epithelium), whole embryos were collected, and Reichert's membrane removed using tungsten needles. Embryos were washed in cold DMEM/F12, then incubated in 0.25% trypsin for 5 min at 37 °C. To dissociate tissue into a single-cell suspension, a 1:1 ratio of DMEM/F12, 20% newborn calf serum, 4 mM EDTA was added. Cell clumps were triturated into single cells by mouth-pipetting using filtered (Sutter Instruments) 75-mm glass capillaries. Single-cell suspensions were filtered through FlowMI cell strainers (40 μm, Sigma-Aldrich) to remove debris. Single cells were spun at 450g for 4 min at room temperature, and cell numbers determined using a Neubauer haemocytometer. For pre-implantation embryo dissociations, embryos were incubated in 0.5% trypsin-EDTA (Invitrogen) at 37 °C for 3 min before transferring to PBS supplemented with 0.5 mM EDTA (Invitrogen) and 4% BSA (Sigma) for mechanical dissociations. Trypsin-treated embryos were dissociated by trituration with pulled capillaries and mouth-pipetting. Dissociated cells were stored in FHM on ice until loading on a Chromium Controller (10x Genomics).

Single-cell library preparation. Cells were counted and diluted to a final concentration in DMEM/F12, 10% fetal bovine serum in Single Cell Master Mix (10x Genomics). Cellular suspensions were loaded on a Chromium Controller⁵⁴ targeting a 2,500–10,000 cell range, depending on tissue type and embryo stage, to generate single-cell 3' RNA-seq libraries, in duplicate or triplicate. Single-cell 3'

RNA-seq libraries were generated following the manufacturer's instructions (10x Genomics Chromium Single Cell 3' Reagent Kit User Guide v2 Chemistry).

Next-generation sequencing of single-cell libraries. Single-cell 3' RNA-seq libraries were quantified on an Agilent Bioanalyzer with a high-sensitivity chip (Agilent), and Kapa DNA quantification kit for Illumina platforms (Roche). Libraries were pooled according to target cell number loaded. To determine the exact number of cells in each library, libraries were sequenced at low depth (2,000 reads per cell) and short reads (40 bp). Sequencing libraries were loaded at 12 pM on an Illumina HiSeq 2500 with 1 × rapid SBS kit v2 (50 cycles) using the following read length: 26-bp read 1, 8-bp I7 index and 40-bp read 2. After sequencing, the number of cells in each single-cell 3' library was calculated using the Cell Ranger analysis pipeline V2.1 (10x Genomics). Library pools were re-made according to the actual number of cells determined in each library for sequencing at a depth of ~200,000 reads per cell and the capacity of an Illumina NovaSeq flow cell. New pools were loaded on an Illumina NovaSeq 6000 using 2 × NovaSeq 6000 S2 reagent kits (200 cycles) and 1 × NovaSeq 6000 S4 reagent kits (300 cycles) using the following read length: 26-bp read 1, 8-bp I7 index and 98-bp read 2.

ISH and immunofluorescence on embryos. For mRNA ISH, post-implantation embryos were fixed in 4% PFA in PBS at 4 °C overnight, then dehydrated through a methanol series and stored at –20 °C. ISH was performed as previously described⁵¹ using antisense riboprobes. Probes used are listed in Supplementary Tables 7, 8.

Immunofluorescence of pre-implantation embryos was performed as previously described^{20,51}. Fixed embryos were washed for 5 min in 0.1% Triton X-100 (Sigma) in PBS (PBX), permeabilized in 0.5% Triton X-100 and 100 mM glycine (Sigma) in PBS for 5 min, washed again in PBX for 5 min and blocked in 2% horse serum (Sigma) in PBS (blocking solution) for 1 h at room temperature before antibody incubation. Embryos were incubated in primary antibodies diluted in blocking solution overnight at 4 °C. Embryos were then washed 3 times for 5 min each in PBX and blocked again for 1 h at room temperature before incubation with secondary antibodies diluted in blocking solution were applied for 1 h 30 min at 4 °C. Embryos were then washed twice for 5 min each in PBX and subsequently incubated with 5 μg/ml Hoechst 33342 (Invitrogen) in PBS to stain nuclei for 5 min or until mounting for imaging.

For immunofluorescence, post-implantation embryos were fixed for 10 min in 4% PFA at room temperature, washed three times in 0.1% Triton-100 in PBS, permeabilized in 0.5% Triton-100 in PBS for 20 min and then washed three times in 0.1% Triton-100 in PBS. Embryos were incubated in blocking buffer containing 2% donkey serum (Jackson Labs) in 0.1% Triton-100 in PBS for 1 h at 4 °C, followed by the incubation in the primary antibodies diluted in blocking buffer overnight at 4 °C. Embryos were washed 3 times in 0.1% Triton-100 in PBS before incubated in secondary antibody overnight at 4 °C, and then washed again 3 times in 0.1% Triton-100 in PBS and counterstained in 5 μg/ml Hoechst. Primary and secondary antibodies used, and their dilutions, are listed in Supplementary Table 9.

Amplification and cloning of antisense riboprobes. Total RNA was isolated from whole 13ss wild-type embryos using Trizol (Invitrogen). Five hundred microlitres of Trizol was added and the sample vortexed. One hundred microlitres of chloroform was added and incubated for 2 min at room temperature. Samples were then centrifuged at 12,000 r.p.m. for 15 min at 4 °C. The aqueous phase was removed, and 1 × volume of isopropanol and 1 μl GlycoBlue coprecipitant (15 mg/ml, Invitrogen) were added to precipitate RNA and visualize the pellet. Samples were incubated at room temperature for 20 min, then centrifuged for 10 min at 12,000 r.p.m. at 4 °C. Samples were placed on ice and washed with 500 μl 75% ethanol and air-dried for 5 min on ice. RNA pellets were resuspended in UltraPure DNase/RNase-free water (Invitrogen). TurboDNase 2 U/μl was used to eliminate DNA, with samples incubated for 30 min at 37 °C. RNA was phenol- and chloroform-extracted and precipitated by adding 1/10 volume 3 M sodium acetate, 2.5 × volumes ethanol at –80 °C for 1 h. Pellets were washed with 75% cold ethanol and air-dried for 5 min on ice, and resuspended in UltraPure DNase/RNase-free water (Invitrogen). Concentrations were determined using a NanoDrop.

cDNA fragments for riboprobes were generated using the SuperScript IV One-Step PCR reverse transcription (RT-PCR) System (Invitrogen) and gene-specific primers (Supplementary Table 8). cDNA fragments were amplified from embryo RNA and cloned into pCR-Blunt II TOPO using the ZeroBlunt TOPO PCR Cloning Kit (Invitrogen). *Rhox5* and *Rhox6* riboprobes were amplified using gene-specific primers (Supplementary Table 8) from RHOX5 and RHOX6 expression vectors, respectively, and cloned into the dual promoter pCR II TOPO vector using the TOPO TA Cloning Kit (Invitrogen). All subcloned ISH probes were validated by sequencing.

Image data acquisition. Wide-field images of ISH of embryos and gut tubes were acquired on a Zeiss AxioZoom stereo-microscope with a Zeiss AxioCam MRC CCD camera and ZEN 2.3 software, using the manual extended depth of focus application which combines sharp regions from several focal planes to produce one resulting image. Laser-scanning confocal images of pre- and post-implantation embryos were acquired on a Zeiss LSM 880. Fixed E5.5 and E6.0 embryos were imaged

in a drop of PBS on a glass-bottom dish (MatTek). Images were acquired using a Plan-Apo 20×/NA0.8 M27 objective. Z-stacks were taken at 0.88- μ m intervals. Pre-implantation embryos were imaged using an EC Plan-Neofluar 40×/NA1.30 oil immersion objective at 1- μ m z-intervals. Fluorescence was excited using a 405-nm diode (Hoechst 33342), 488-nm argon, 561-nm DPSS-561-10 and HeNe 633-nm lasers. Raw data were processed in ZEN (Zeiss, <https://www.zeiss.com/microscopy/us/products/microscope-software/zen.html>) or Imaris (Bitplane, <http://www.bitplane.com/>) software, and assembled in Adobe Photoshop or Illustrator (Adobe Creative Cloud, <https://www.adobe.com/creativecloud.html>).

Image data analysis and processing. Three-dimensional reconstructions of the distribution of nuclei of definitive and visceral endoderm descendants within gut tube was performed using Neurolucida software (<https://www.mbfioscience.com/neurolucida>). Serial transverse sections of three 13ss *Afp-GFP^{GTG+}* embryos were cut and counterstained with Hoechst 33342 (10 mg/ml, Invitrogen) to label nuclei. Sections were imaged on an AxioImager M1 (Zeiss) using a Hamamatsu C10600 Orca-R21 camera. The outline of the gut tube was traced on each section at low magnification (5×/NA0.16 objective), then nuclei of all cells (GFP-positive and GFP-negative) and visceral endoderm descendants (GFP-positive) were counted at high magnification (40×/NA 0.75 objective). Nuclei identified in serial sections were used to reconstruct a 3D image that depicts the distribution of definitive and visceral endoderm descendants along entire gut tubes.

Fluorescence-activated cell sorting. Single cells recovered from E7.5 endoderm (comprising visceral and definitive endoderm), as well as E8.75 gut tube, parietal endoderm and yolk sac endoderm were resuspended in serum-free DMEM/F12 medium and sorted before scRNA-seq using a SORP FACSARIA Iiu (BD Biosciences), with a 100- μ m nozzle at 137.9 kPa (20 psi) in purity mode. Cell suspensions were sorted based on GFP content, with both GFP-positive and GFP-negative fractions collected, and dead cells excluded using ethidium homodimer-1 (Ethd-1, 4 μ M, Thermo Fisher). Debris was excluded from yolk sac endoderm and parietal endoderm cell suspensions by selecting calcein violet (0.05 μ M, Thermo Fisher) and excluding Ethd-1-positive events. GFP, calcein violet and Ethd-1 were excited at 488, 561 and 405 nm respectively, and detected using 530/50-, 582/15- and 450/50-nm band-pass filters, respectively. Sorted cells were collected in DMEM/F12, 10% newborn calf serum, resuspended immediately after sorting in collection buffer, and counted before loading on a 10x Chromium Controller. Wherever possible, purity checks were performed indicating >99.9% sample purity. Gating strategies for each tissue collected are provided in Supplementary Fig. 1.

RNA isolation and next-generation sequencing of bulk tissue. Total RNA was extracted from bulk tissue and pooled dissociated cells of 13ss (~E8.75) gut tubes, from bulk tissue (gut tube quarters) representing anterior, anterior midgut, midgut posterior and posterior sections of 13ss gut tubes (Extended Data Fig. 1), as well as from extra-embryonic visceral endoderm and embryonic visceral endoderm at E7.5. The Trizol method (Invitrogen) was used for RNA extraction. RNA concentration and quality were assessed, and cDNA libraries construction and sequencing were performed by the Genomics and Epigenomics Core Facility at Weill Medical College (Cornell University). Paired-end sequencing (Illumina HiSeq 4000, 50-bp reads) was performed.

Bulk RNA-seq processing. The bulk RNA-seq expression datasets generated are listed in Supplementary Table 13. All samples were generated in duplicate. Bulk RNA-seq data were aligned to the mm38 mouse genome using STAR⁵⁵ and reads that mapped to multiple genomic locations were filtered out. Gene expression counts for each sample were determined using the summarizeOverlaps function of the GenomicRanges package using Ensembl annotations⁵⁶. The annotations and STAR parameters used for scRNA-seq data alignment were also used for bulk RNA-seq data to maintain consistency.

Bulk RNA-seq and scRNA-seq data were compared by computing the Pearson's correlation between log-transformed bulk counts and aggregated molecule counts across all relevant single cells (Extended Data Fig. 2b). DESeq2⁵⁷ was used to determine the differentially expressed genes between E7.5 exVE and emVE tissues (Extended Data Fig. 7d). DESeq2 was also used to normalize the bulk data for determining spatial patterns of gene expression in the E8.75 gut tube (Extended Data Fig. 8c).

Palantir. Alignment of cells along their developmental trajectories was performed using Palantir, a recently published trajectory-detection algorithm^{10,11}. A key distinguishing feature of Palantir is that—rather than treating lineage decisions as bifurcations—cell fate choices are modelled as continuous probabilistic processes. Palantir accomplishes this by estimating the probability of a cell in an intermediate state to reach any of the terminal states. The entropy of these branch probabilities has been shown to represent a quantitative measure of the differentiation potential or plasticity of the cell, in which multipotent cells have the highest differentiation potential and mature terminal states have the lowest potential. The high resolution achieved by Palantir allows detailed mapping of gene expression trends and dynamics that are correlated with changes in lineage potential^{10,11}. See the Supplementary Note 7 on Palantir for details of the interpretation of Palantir results and visualization.

The different parameters used for Palantir are listed in Supplementary Table 14. Harmony (Supplementary Note 1) was used to compute the augmented affinity matrix by determining the mutually nearest neighbours between successive time points. Diffusion components were computed by using the Harmony augmented affinity matrix and used as inputs for Palantir.

A $k = 30$ value was used for datasets that involved pre-implantation stages, because the number of cells is relatively lower in these stages. The number of neighbours was increased for datasets with increasing complexity. The number of diffusion components was chosen on the basis of the Eigen gap for each dataset. Palantir results, however, have previously been shown to be robust to these parameters^{10,11}.

Gene-expression trends were determined as described in Palantir using the branch probabilities and generalized additive models. Similarly, the clustering of the trends was performed as described in Palantir, using Phenograph¹⁴.

Trophectoderm lineage decision. Harmony was used to generate an augmented affinity matrix that spans cells of all lineages (ICM, EPI, PrE and trophoctoderm) across E3.5 and E4.5. The cells were projected onto diffusion components using this affinity matrix with the number of components (two components) chosen by Eigen gap. The distance between any two cells is measured using the multiscale distance (see 'Multiscale distance' in Supplementary Note 1).

The average distance between pairs of ICM and trophoctoderm cells (13.9) at E3.5 is orders of magnitude greater than the distance between pairs of ICM and EPI (0.41) or PrE (2.0) cells (Fig. 2b). This suggests that the lineage decision between ICM at E3.5 and trophoctoderm occurs at a stage earlier than E3.5.

Relationship between EPI, visceral endoderm and ExE cells at E5.5. E5.5 EPI, visceral endoderm and ExE cells were projected onto a low-dimensional embedding using diffusion maps. The number of components (10) for the embedding was chosen by Eigen gap among the top diffusion components. Similar to the analysis above, the ExE cells at E5.5 continue to be substantially further from EPI and visceral endoderm cells (Extended Data Fig. 6a, b).

To test the relationships between the E5.5 cell lineages, we first identified the cells that form the boundaries for the different lineages by identifying the extremes of the diffusion components for each cell lineage (Extended Data Fig. 6c). We then constructed a kNN graph in the embedded space and computed the shortest paths between the EPI boundary to the visceral endoderm and ExE boundary cells. The path from EPI to ExE boundary cells includes steps that are substantially distant (Extended Data Fig. 6d). By contrast, the path from EPI to visceral endoderm boundary cells includes relatively uniform step sizes and does not include large steps, which indicates continuity (Extended Data Fig. 6d).

Identification of E5.5 emVE and exVE cells. The visceral endoderm trajectories using cells from E3.5 to E8.5 show the following properties: (1) E3.5 and E4.5 cells do not show any change in differentiation potential along pseudo-time (Supplementary Fig. 5a) and thus are representative of uncommitted cells, (2) emVE cells at E6.5 and E7.5 show an increasing probability towards the gut tube, and (3) exVE cells at E6.5 and E7.5 show an increasing probability towards the yolk sac endoderm (Extended Data Fig. 7e, Supplementary Fig. 5a). We therefore used the Palantir branch probabilities to identify putative E5.5 uncommitted, emVE and exVE cells that follow these properties: (1) E5.5 cells with the same differentiation potential as E3.5 and E4.5 cells are designated uncommitted cells, and (2) E5.5 cells with gut tube and yolk sac probabilities greater than E3.5 probabilities (epsilon 0.01) are designated putative emVE and exVE cells, respectively (Extended Data Fig. 7e, f).

emVE and exVE gene signature. Covariance matrices were computed separately for the putative E5.5 emVE (449 cells) and exVE cells (618 cells) using the 2,500 most-variable genes, which were also used for characterizing visceral endoderm developmental trajectories with Palantir (Fig. 4a). MAGIC imputed data were used for computing the covariances. Hierarchical clustering was used to identify clusters of covarying genes in each of these compartments. Visual inspection revealed the presence of two clusters of genes in emVE with strong intra-cluster correlation, but anti correlated across the clusters (Supplementary Fig. 5b). The genes comprising these clusters were identified by cutting the hierarchical clustering tree to yield three clusters (Supplementary Fig. 5b).

Additional information. Additional information regarding methods can be found in the Supplementary Information. For scRNA-seq data processing, see Supplementary Note 1; for the Harmony framework, see Supplementary Note 2; for the manifold classifier, see Supplementary Note 3; for anterior–posterior pseudo-space ordering, see Supplementary Note 4; for the visceral and definitive endoderm classifier, see Supplementary Note 5; for the identification of transcription factors that are predictive of anterior–posterior pseudo-space in the gut tube, see Supplementary Note 6; and for an overview of the Palantir algorithm and use of differentiation potential and branch probabilities to infer lineage decisions, see Supplementary Note 7.

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

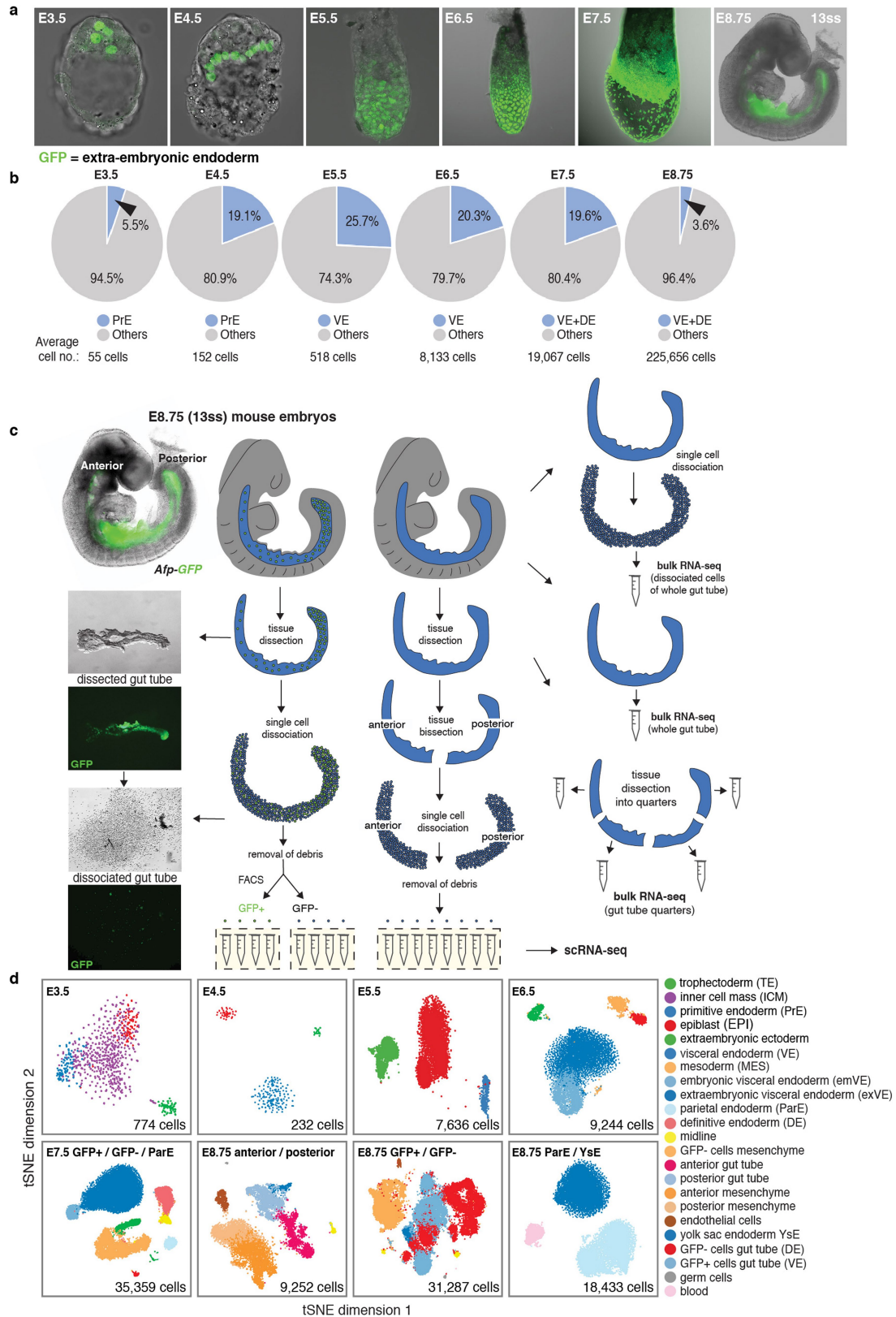
Data availability

All the generated data including bulk and scRNA-seq data are available through the Gene Expression Omnibus, under accession numbers GSE123046 (scRNA-seq) and GSE123124 (bulk RNA-seq). The data can be explored at <https://endoderm-explorer.com>, and any other relevant data are available from the corresponding authors upon reasonable request.

Code availability

Harmony is available as a Python module at <https://github.com/dpeerlab/Harmony>. Palantir is available as a Python module at <https://github.com/dpeerlab/Palantir>. A Jupyter notebook detailing the usage of Harmony along with sample data is available at http://nbviewer.jupyter.org/github/dpeerlab/Harmony/blob/master/notebooks/Harmony_sample_notebook.ipynb.

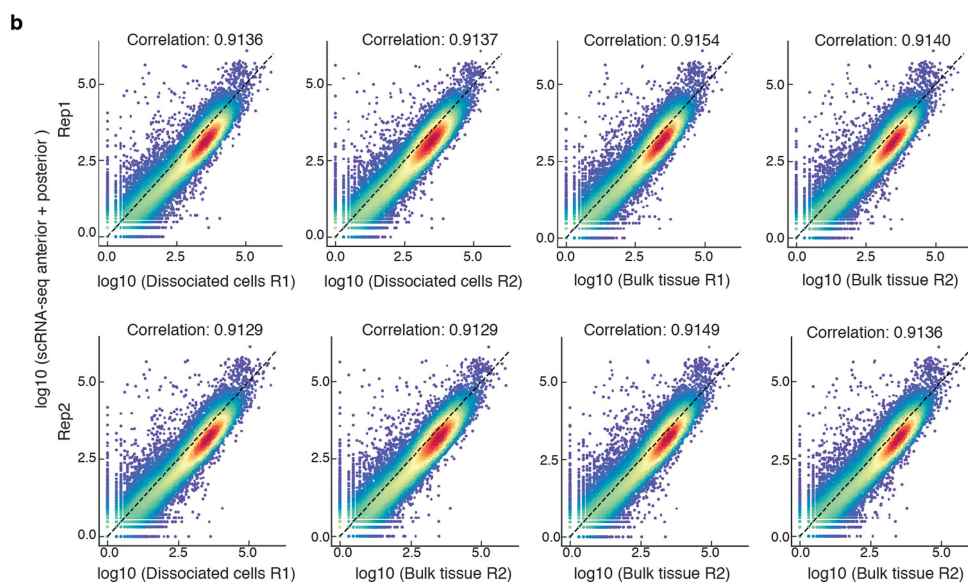
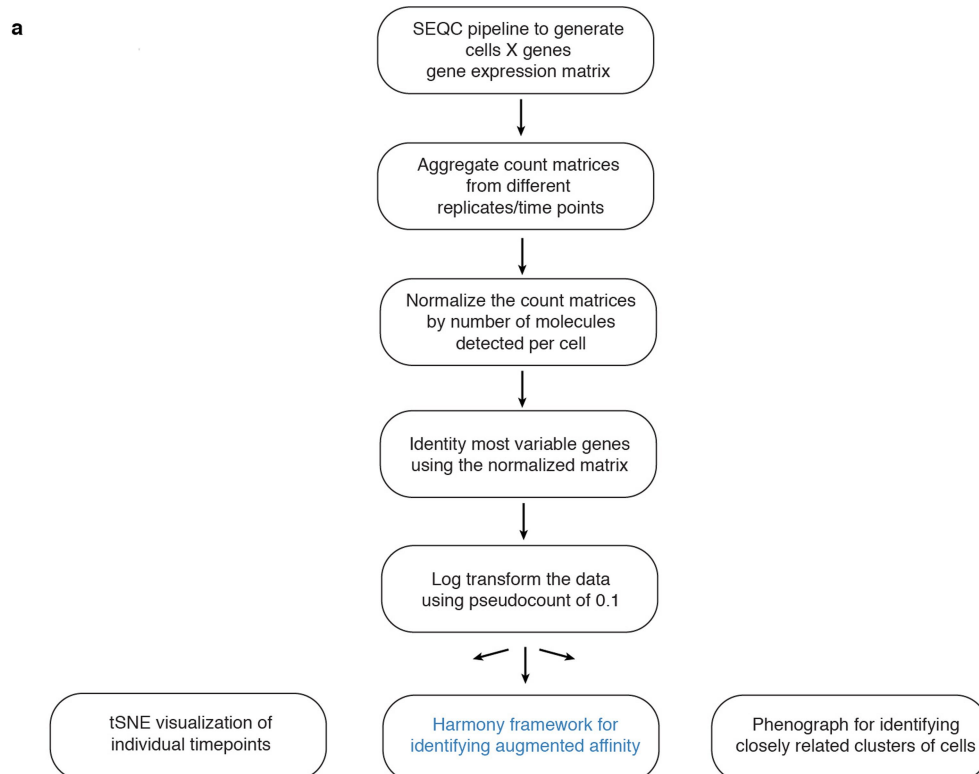
51. Behringer, R. G. M., Nagy, K. V. and Nagy, A. *Manipulating the Mouse Embryo: A Laboratory Manual*, 4th edn (Cold Spring Harbor Laboratory, Cold Spring Harbor, 2014).
52. Downs, K. M. & Davies, T. Staging of gastrulating mouse embryos by morphological landmarks in the dissecting microscope. *Development* **118**, 1255–1266 (1993).
53. Morgani, S. M. et al. Totipotent embryonic stem cells arise in ground-state culture conditions. *Cell Reports* **3**, 1945–1957 (2013).
54. Zheng, G. H. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
55. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
56. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
57. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).



Extended Data Fig. 1 | See next page for caption.

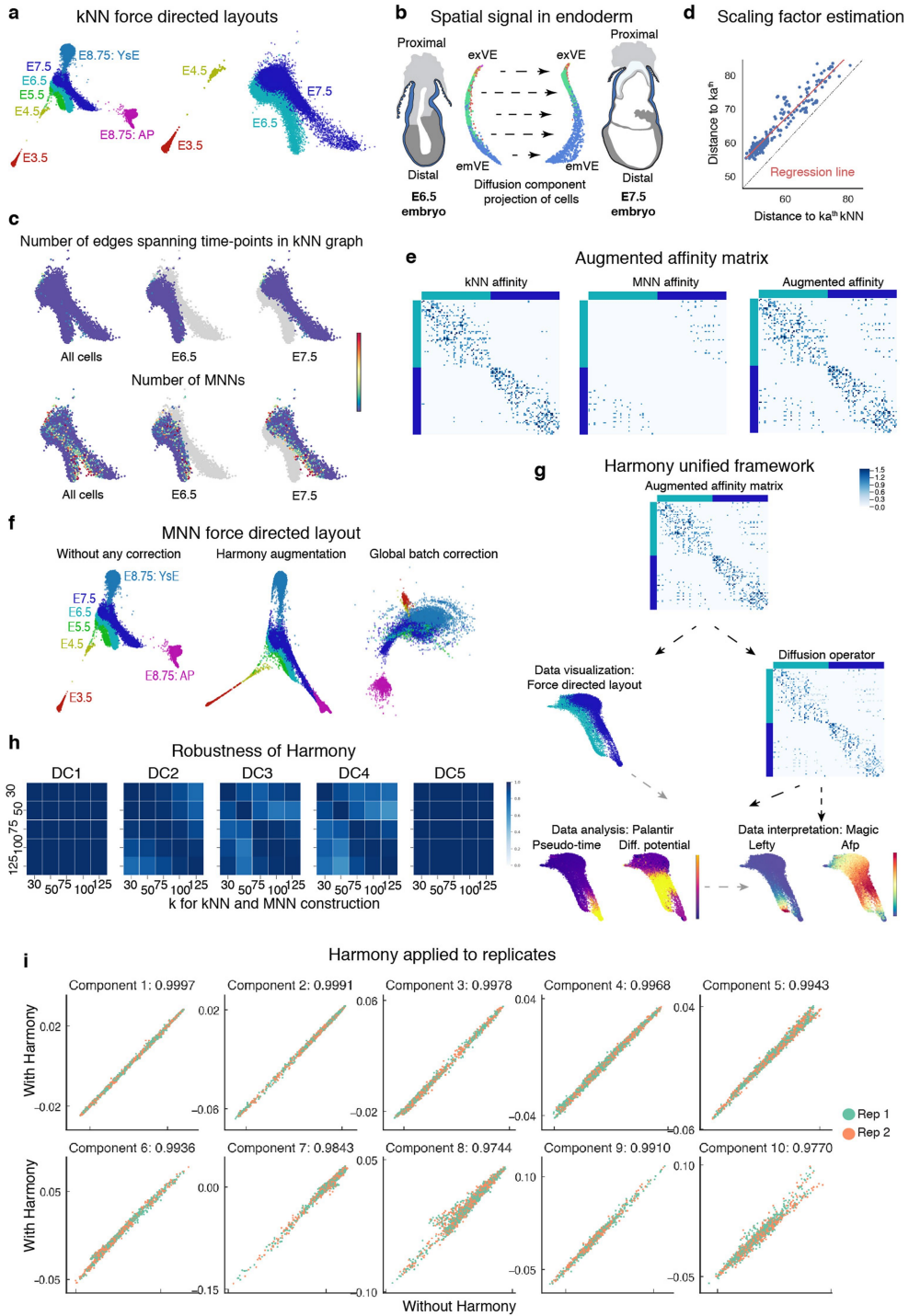
Extended Data Fig. 1 | Endoderm cell representation in mouse embryos, from blastocyst through midgestation, and single-cell collection pipeline. **a**, Distribution of extra-embryonic endoderm cells (GFP, green) from blastocyst (E3.5) to midgestation (E8.75, 13ss) demarcated using *Pdgfra*^{H2B-GFP20} (pre-implantation stages) and *Afp-GFP*¹² (post-implantation stages) reporters. Extra-embryonic endoderm (PrE and visceral endoderm derivatives) cells contribute to the gut tube of the E8.75 embryo. **b**, Pie charts depicting the fraction of endoderm cells per embryo, for all stages analysed in this study. **c**, Schematic of protocol used for single-cell collection, with E8.75 gut tube provided as an example. Gut tubes were micro-dissected from embryos, then dissociated into single

cells. Single cells of either anterior and posterior halves of gut tubes, or AFP-GFP-positive (visceral endoderm descendants) and AFP-GFP-negative (definitive endoderm descendants) collected using fluorescence-activated cell sorting, were used for single-cell 3' mRNA library construction on the 10x Genomics Chromium platform. For bulk RNA-seq, whole gut tubes that had been dissociated into single cells and then pooled, whole intact gut tubes and whole gut tubes dissected into quarters were collected for sequencing. **d**, *t*-SNE plots of collected libraries for each time point with each dot representing a single cell. Phenograph was used to identify clusters of cells, colour-coded by cell type with annotation based on expression of known markers.



Extended Data Fig. 2 | Computational pipeline and comparison of scRNA-seq data with bulk RNA-seq data. **a**, Flow chart of computational data processing pipeline. **b**, Plots showing the Pearson's correlation

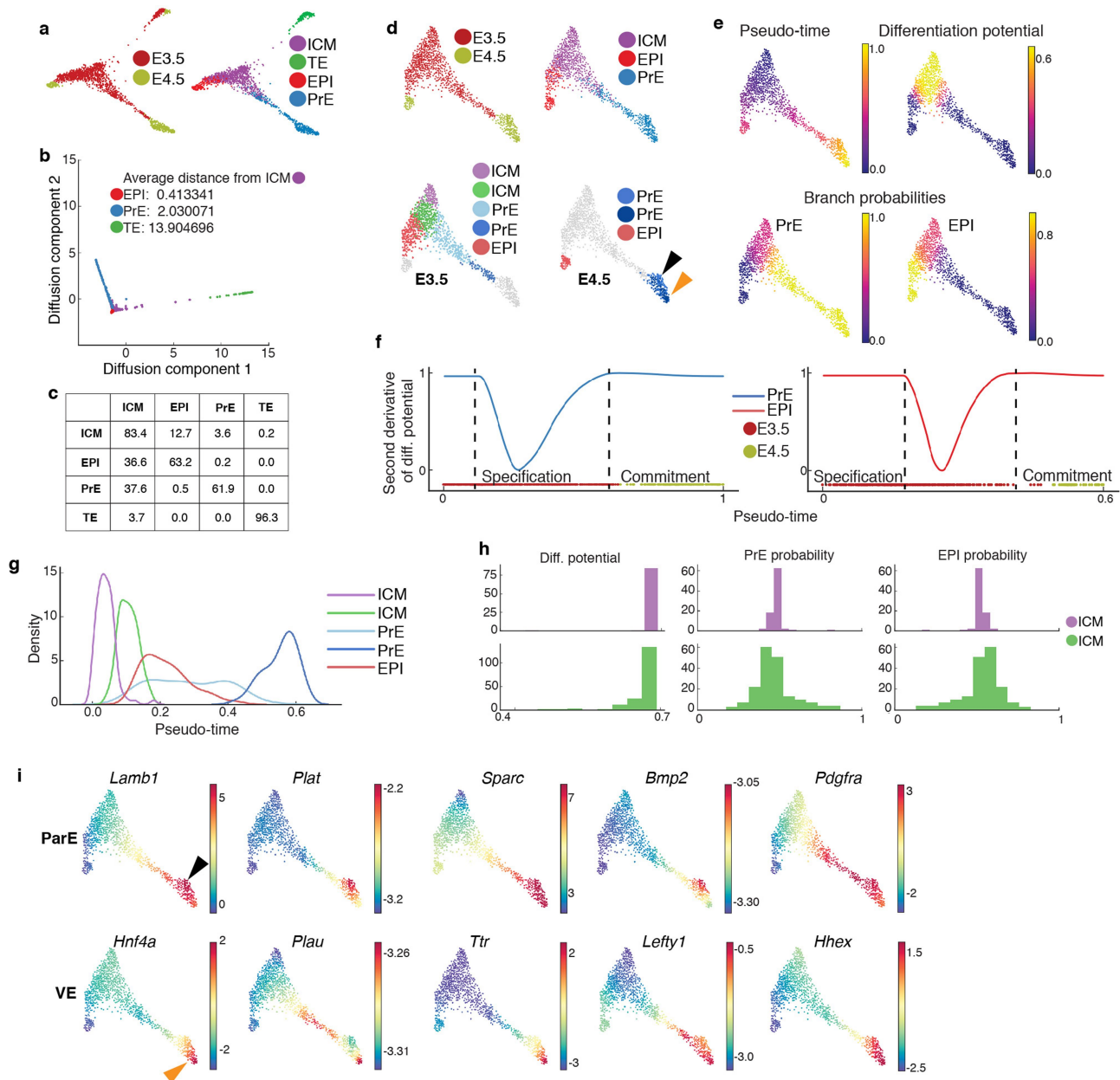
between aggregated scRNA-seq data of anterior and posterior halves of the gut tube with bulk RNA-seq of dissociated (and pooled) cells and bulk tissue, respectively. The two rows represent two replicates.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Mutually nearest neighbour augmentation to correct batch effects between time points, and Harmony unified framework for scRNA-seq data analysis. **a**, Force-directed layouts for cells of the E3.5, E4.5, E5.5, E6.5, E7.5 and E8.75 (amalgamation of anterior and posterior gut tube halves) time points. Cells are coloured by time point. The graph was generated using an adjacency matrix derived from the standard k NN graph. Differences between consecutive time points represent underlying developmental changes, but are also confounded by technical batch effects—including discontinuity between E3.5 and E4.5, and the lack of spatial alignment between E6.5 and E7.5. **b**, E6.5 and E7.5 cells projected along their respective first two diffusion components. These projections reveal a dominant first component with strong spatial signal within individual time points. Cells are coloured by Phenograph clusters. **c**, Top, the number of edges connecting cells between time points are limited in the k NN graph. Bottom, plots showing the number of mutually nearest neighbours (MNNs) between E6.5 and E7.5 time points. The MNNs are enriched along the boundary between time points, providing support for augmentation of the k NN graph with additional edges between MNNs between the consecutive time points. **d**, The MNN distances can be converted to affinities on a similar scale as the k NN affinities, using linear regression to determine the relationship between the two scales, k NN and ($l < k$) MNN distances. **e**, Example of

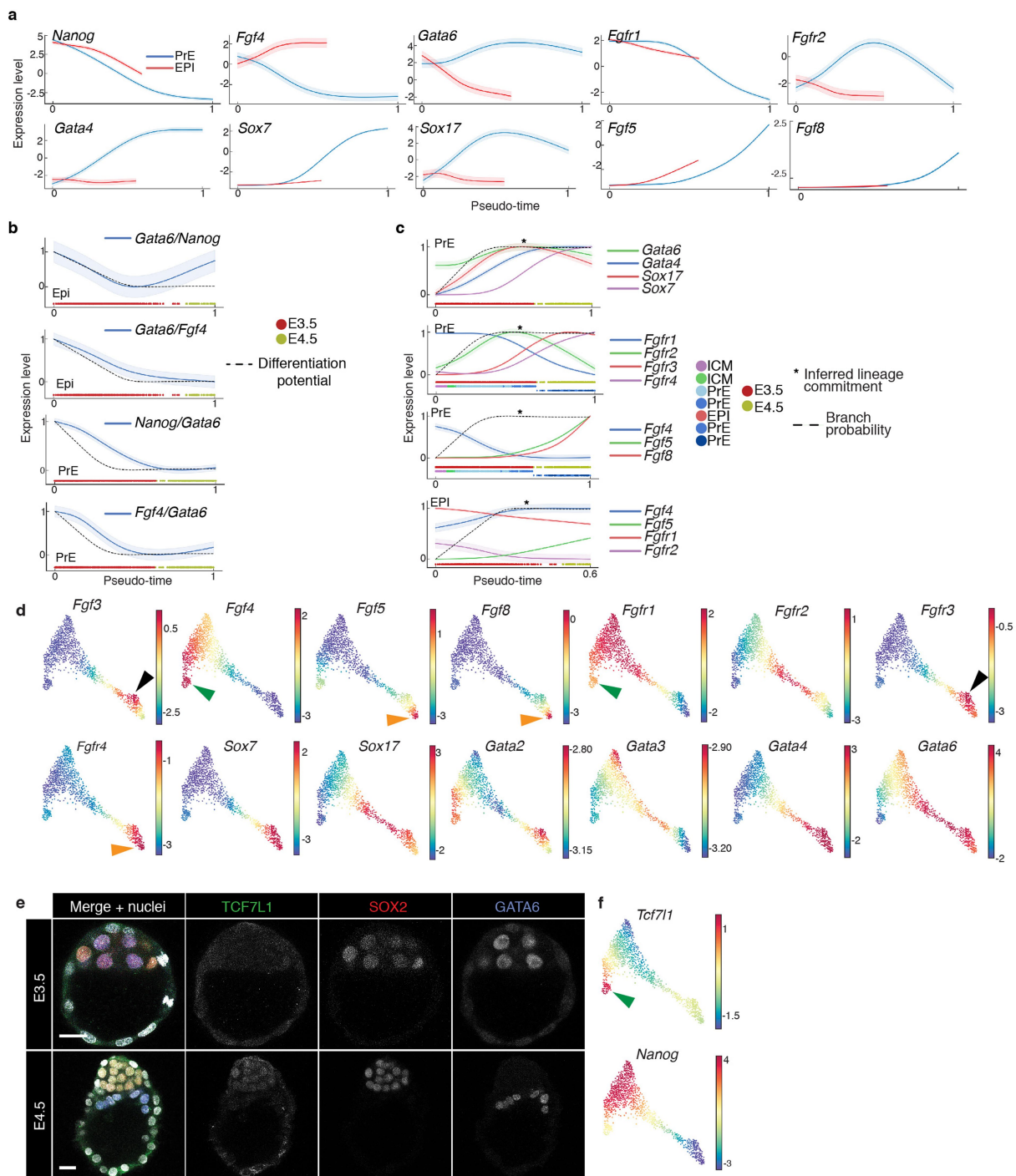
the augmented MNN affinity matrix construction. Left, k NN affinities for a subset of E6.5 and E7.5 cells. Middle, MNN affinity matrix determined using linear regression (in **d**) to convert distances E6.5 and E7.5 cells to affinities. Right, augmented affinity matrix: sum of the k NN and MNN affinity matrices. **f**, Comparison of force-directed layouts. Left, standard k NN affinity matrix. Middle, Harmony augmented affinity matrix. Right, plot generated using `mnnCorrect`¹⁸ for global batch effect correction, leading to ‘over-correction’ and loss in signal between time points. **g**, Harmony framework starts with the augmented affinity matrix generated as described in Supplementary Note 2. The augmented affinity matrix is used to generate the force-directed graph for visualizing the data. The same augmented matrix is used to compute the diffusion operator for determining the diffusion components, which forms the basis for (1) Palantir trajectory detection and (2) MAGIC imputation. **h**, Robustness of Harmony. Plots showing the correlation between diffusion components for different values of k , the number of nearest neighbours for k NN graph construction. Visceral endoderm cells in Fig. 4 were used for testing robustness. **i**, Harmony applied to replicates. Plots showing the Pearson’s correlation between diffusion components without Harmony (x axis) and with Harmony applied between the two replicates of the E8.75 gut tube. Plots shown for 3,512 cells.



Extended Data Fig. 4 | Lineage decisions in the mammalian blastocyst.

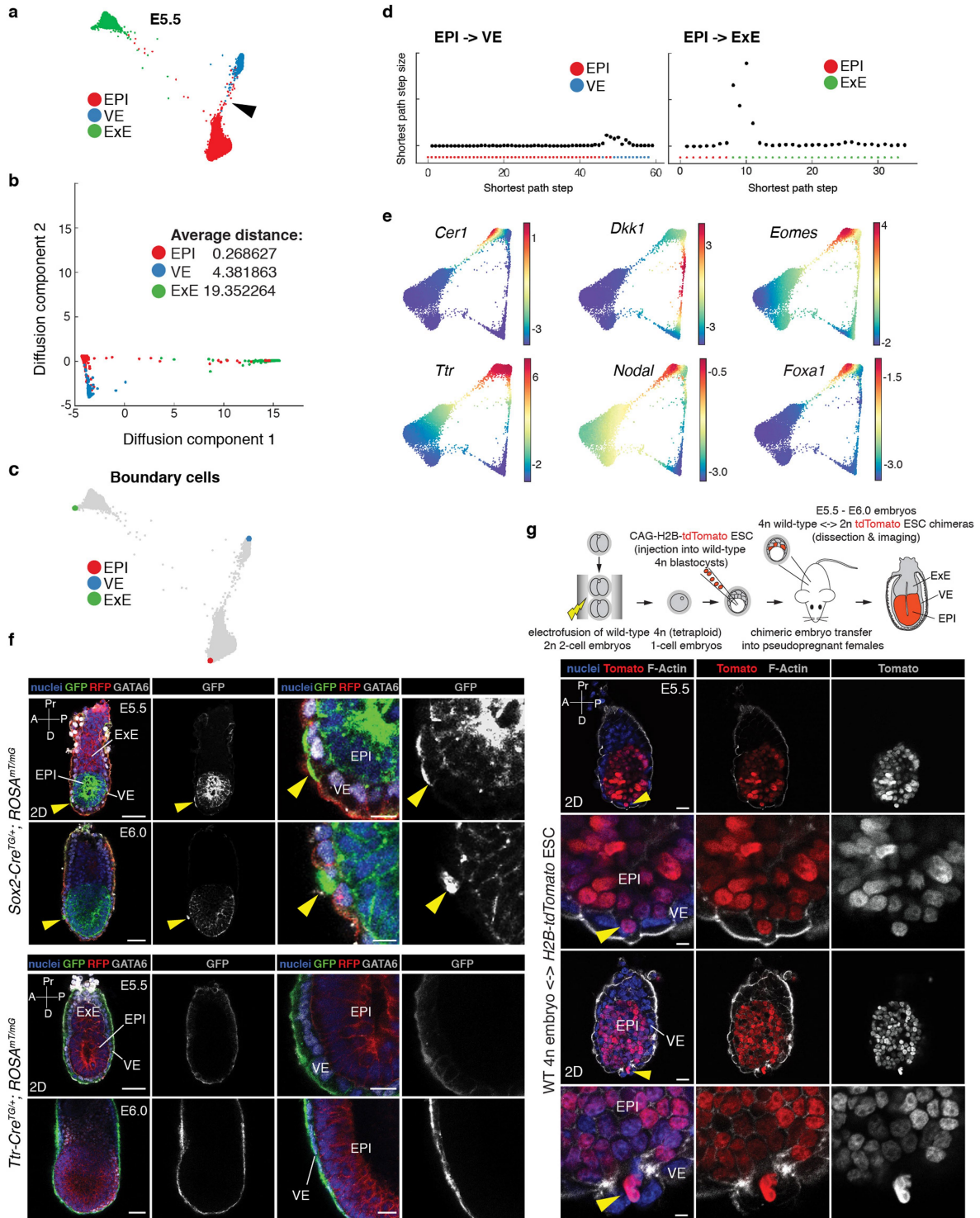
Results from pooling cells of two replicates at E3.5 and E4.5, followed by Harmony augmentation. **a**, Force-directed layout of E3.5 and E4.5 cells depicting the relationship between three blastocyst lineages. Cells coloured by time point or annotated cell types. **b**, Plot showing projection of E3.5 and E4.5 cells along the first two diffusion components. Distances between lineages were computed using multiscale distances. **c**, Table showing the connectivity between different compartments in a k NN graph of E3.5 cells. Each row represents the fraction of outgoing edges from cells of the respective compartment that connect to cells in the compartments specified in the columns. **d**, Force-directed layout of E3.5 and E4.5 cells after removal of trophoblast cells, showing relationships between ICM, EPI and PrE cells. Cells are coloured by time point or Phenograph¹⁴ clusters. **e**, Palantir^{10,11} determined pseudo-time

ordering, differentiation potential and branch probabilities of PrE and EPI cell lineages. **f**, Plots showing the second derivative of PrE and EPI differentiation potential along pseudo-time, which suggests that changes in differentiation potential—and therefore lineage commitment—in both lineages occur at E3.5. Points of highest changes along pseudo-time represent inferred lineage specification and commitment. **g**, Distribution of E3.5 lineage cells along pseudo-time. Each distribution represents cells from one Phenograph cluster. **h**, Histograms showing the distribution of differentiation potential (left), PrE fate probability (middle) and EPI fate probability (right) in the E3.5 ICM clusters. **i**, Gene expression patterns of parietal and visceral endoderm markers. Each cell is coloured on the basis of its MAGIC⁵⁰ imputed expression level for the indicated gene. Black and orange arrowheads mark presumptive parietal and visceral endoderm lineages, respectively.



Extended Data Fig. 5 | Gene expression trends in EPI, PrE, visceral and parietal enderm lineages in the blastocyst. a, Plots comparing gene expression trends along pseudo-time for genes that encode components of the FGF signalling pathway (*Fgf4*, *Fgf5*, *Fgf8*, *Fgfr1*, *Fgfr2* and *Spry4*), the endoderm-marker transcription factors *Gata6*, *Gata4*, *Sox7* and *Sox17*, and *Nanog* during EPI and PrE lineage specification. Solid line represents the mean expression trend and shaded region represents ± 1 s.d. **b,** Dynamics of transcription factor ratios as lineages emerge. Ratio between *Gata6* and *Nanog*, and *Gata6* and *Fgf4* along EPI; and between *Nanog* and *Gata6*, and *Fgf4* and *Gata6* along PrE, compared to changes in differentiation potential (dotted line). Transcription factor ratios were computed for each cell by using the MAGIC⁵⁰ imputed data for each gene. **c,** Plots comparing gene expression trends along pseudo-

time. Gata and Sox transcription factors, Fgf receptors or Fgf ligands during PrE or EPI specification. Colours at the bottom of each panel represent time point and—where applicable—E3.5 and E4.5 Phenograph clusters. Dashed lines represent branch probabilities in commitment towards respective lineages. **d,** Gene expression patterns of FGF signalling pathway components, and Gata and Sox transcription factor genes. Orange, black and green arrowheads point to high expression in parietal endoderm, visceral endoderm and EPI, respectively. **e,** Laser-scanning confocal data depicting TCF7L1 expression at E3.5 (top panel) ($n = 14$) and E4.5 (bottom panel) ($n = 11$). SOX2 and GATA6 were used as EPI and PrE lineage markers, respectively. **f,** Gene expression patterns of *Tcf7l1* and *Nanog* depicting similar expression of *Tcf7l1* in EPI as *Fgf4* (green arrowhead).

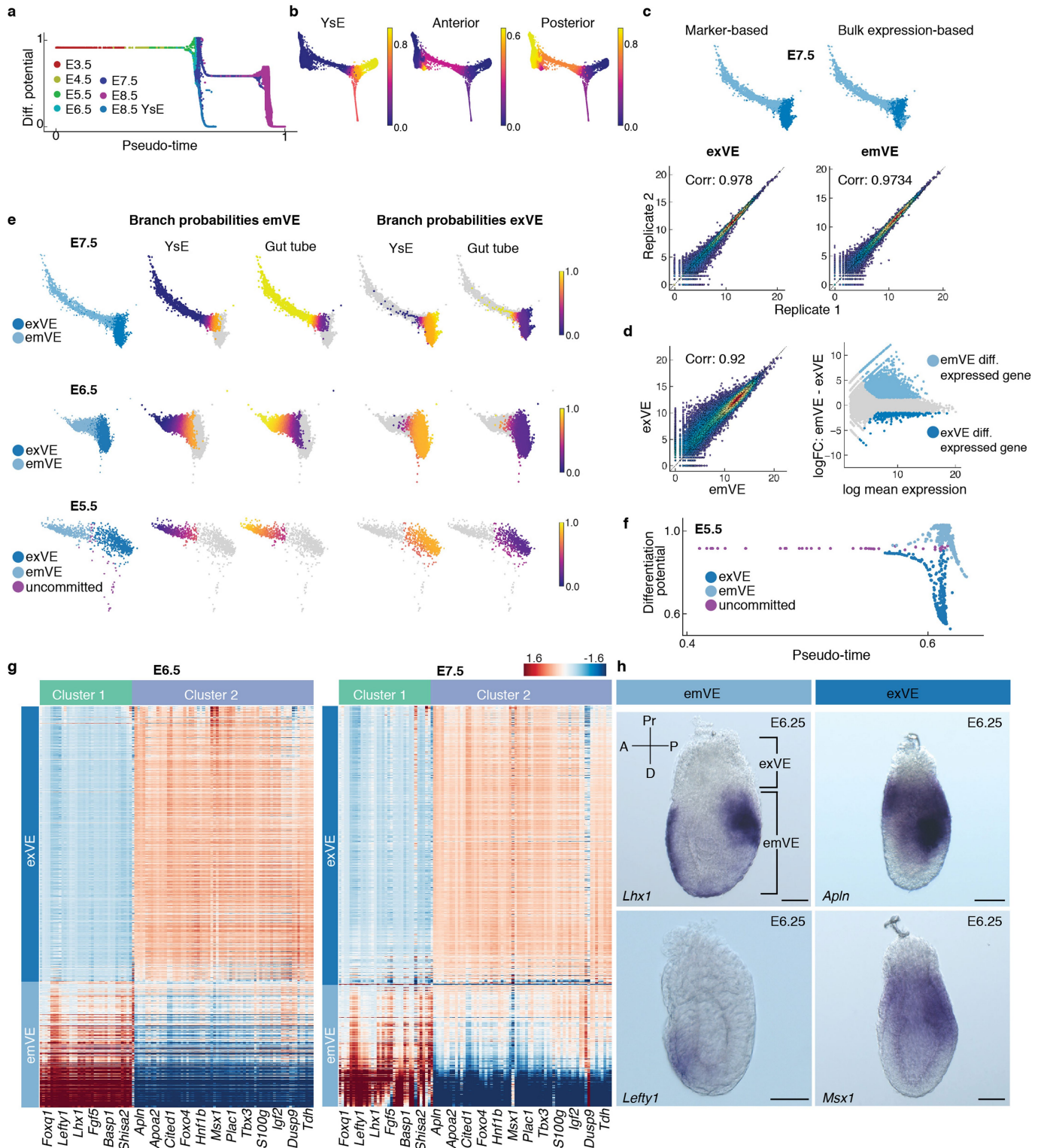


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Force-directed layouts of single E5.5 cells reveal relationships between EPI, visceral endoderm and ExE lineages.

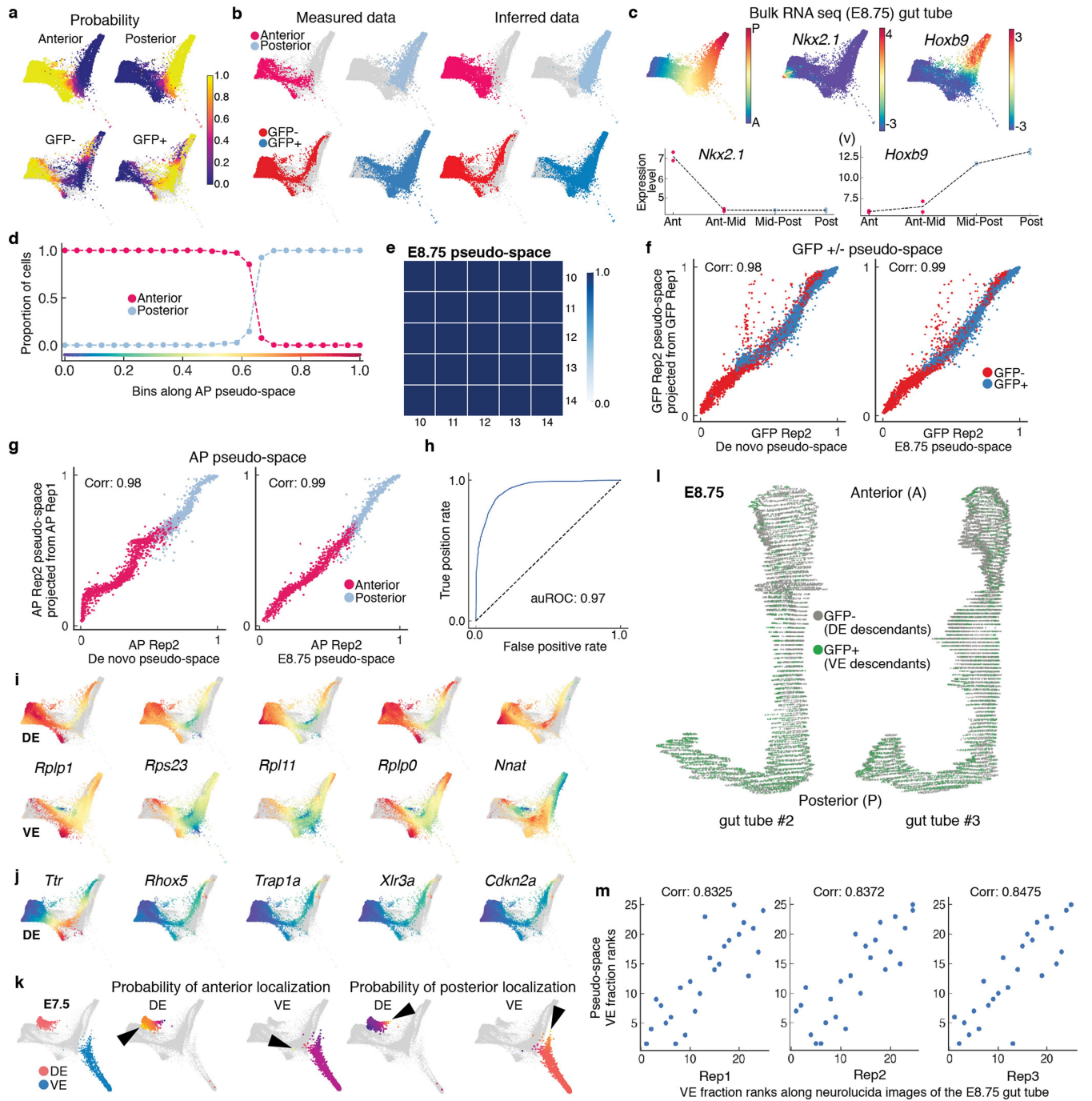
a, Force-directed layouts of E5.5 data generated after pooling replicates, showing the relationship between EPI, visceral endoderm and ExE lineages. Cells are coloured by cell type. Black arrowheads mark cells that transdifferentiate from EPI to visceral endoderm. **b**, Plot showing the projection of EPI, visceral endoderm and ExE cells along the first two diffusion components. Distances between lineages were computed using multiscale distances. **c**, Plots highlighting the extremes of the diffusion components, serving as the boundaries of the phenotypic space for each lineage identity. **d**, Plots showing the shortest path-step sizes for paths from EPI to visceral endoderm (left) and EPI to ExE (right). **e**, Gene expression plots of AVE (*Cer1* and *Dkk1*), visceral endoderm (*Eomes*, *Foxa1* and *Ttr*), and visceral endoderm and EPI (*Nodal*) markers along EPI, and PrE and visceral endoderm lineages from E3.5–E5.5. Cells coloured on the basis of marker expression of indicated gene after MAGIC⁵⁰. **f**, Laser-scanning confocal images of E5.5 and E6.0 *Sox2-cre*^{TG/+};*Rosa26*^{mT/mG} and *Ttr-cre*^{TG/+};*Rosa26*^{mT/mG} embryos

immunostained for GFP, RFP (red fluorescent protein, membrane-localized tdTomato) and GATA6 (a marker of endoderm identity). Cell nuclei stained with Hoechst, and membranes labelled with RFP. Yellow arrowheads point to cells of EPI origin that are present within the visceral endoderm epithelial layer ($n = 10/20$ GFP-positive cells in visceral endoderm of *Sox2-cre*^{TG/+};*Rosa26*^{mT/mG} embryos; $n = 0/27$ GFP-positive cells in the EPI of *Ttr-cre*^{TG/+};*Rosa26*^{mT/mG} embryos). Results validated in at least three independent experiments. Scale bars, 50 μm (low-magnification images), 20 μm (high-magnification images). **g**, Laser-scanning confocal images of two E5.5 wild-type $4n$ embryo \leftrightarrow H2B-tdTomato embryonic stem cell (ESC) chimaeras. Top two rows and bottom two rows represent low- and high-magnification 2D images, respectively ($n = 9/19$ embryo chimaeras showed tdTomato-positive cells in the visceral endoderm). Yellow arrowheads indicate EPI cells intercalating into the visceral endoderm layer. Embryos are counterstained with Hoechst to label nuclei, and phalloidin to label F-actin. Scale bars, 20 μm (low-magnification images), 10 μm (high-magnification images).



Extended Data Fig. 7 | Emergence of spatial patterning of the embryo at E5.5. **a**, Plot showing Palantir pseudo-time versus differentiation potential of visceral endoderm cells from stage E3.5 to E8.75. Drops in differential potential occur at two time points. The first occurs at E5.5, as cells acquire a distal versus proximal fate; the second occurs at E7.5, as cells acquire an anterior versus posterior fate. **b**, Plots of branch probabilities of commitment towards yolk sac endoderm, and anterior and posterior gut endoderm. **c**, Marker based (top left) and bulk RNA-seq based (top right) prediction of exVE and emVE at E7.5. Bottom, plots show the Pearson's correlation between bulk RNA-seq replicates of exVE (291 genes) and emVE (2,239 genes) derived using bulk RNA-seq data. **e**, Plots showing the branch probabilities of E7.5, E6.5 and E5.5 exVE and

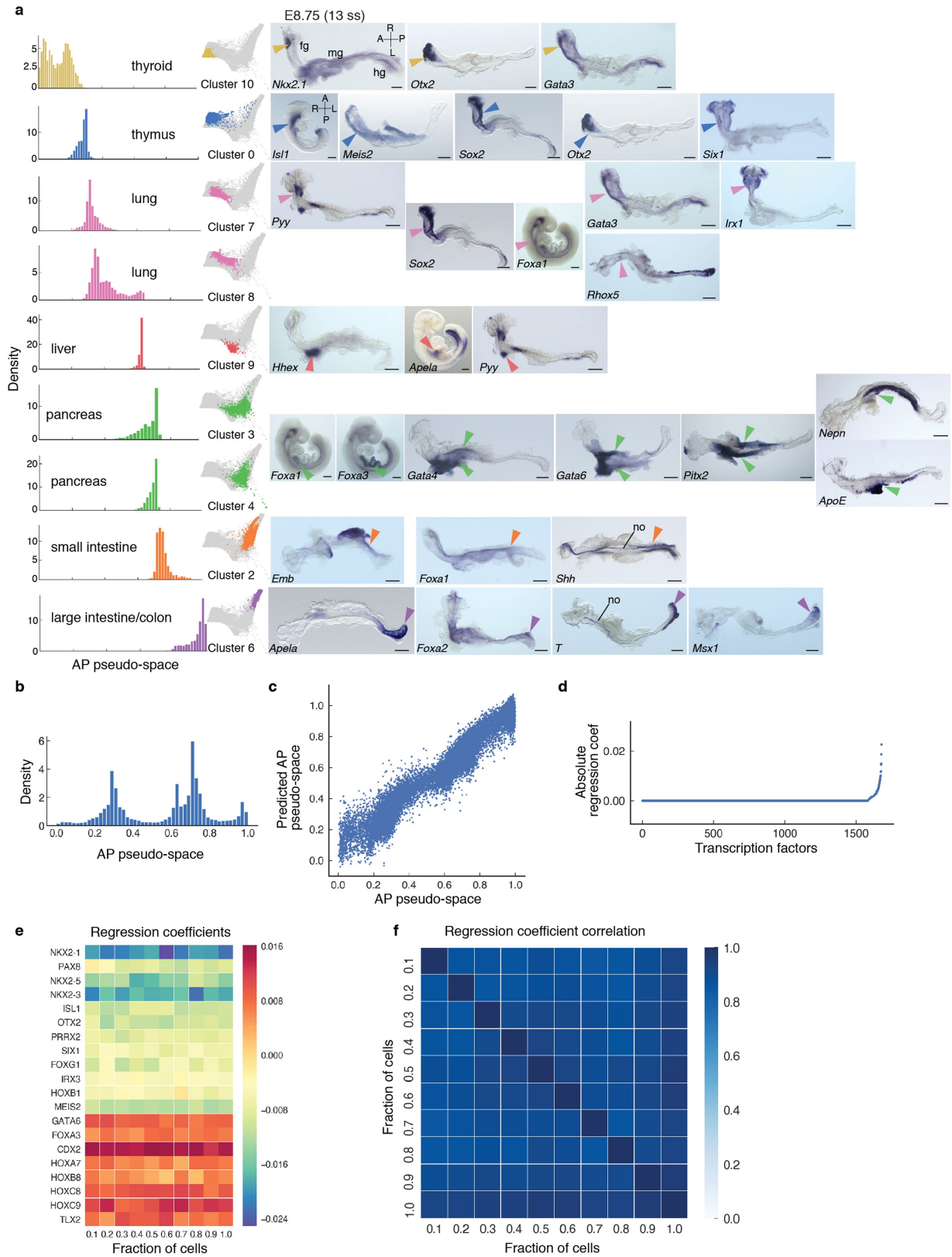
emVE cells to commit towards yolk sac endoderm (extra-embryonic) and gut tube (embryonic). Cells are labeled as exVE and emVE on the basis of expression of known markers (leftmost plot), match expected Palantir branch probabilities (four plots on the right). The branch probabilities of E5.5 cells in committing towards yolk sac endoderm and gut tube were used to infer putative exVE and emVE identities at E5.5. **f**, Plot showing pseudo-time versus differentiation potential of endoderm cells at E5.5, coloured by the inferred cell type. This panel is a magnified view of **a**. **g**, Heat maps of highly expressed genes specifically in exVE or emVE at E5.5 also distinguish exVE and emVE cells at E6.5 and E7.5. **h**, ISH of E6.25 embryos showing expression of *Lhx1* ($n = 3$) and *Lefty1* ($n = 3$), which are genes that are specific for emVE, and *Apln* ($n = 3$) and *Msx1* ($n = 3$), which are genes that are specific for exVE. Scale bars, 50 μm .



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Characterization of E8.75 gut tube anterior–posterior pseudo-space. **a**, Force-directed layout as in Fig. 5. Top, plots show the probabilities of anterior–posterior positioning for the AFP–GFP-positive and AFP–GFP-negative cells inferred using the manifold classifier trained on anterior–posterior cells. Bottom, plots show the probabilities of GFP-positive and GFP-negative status for the cells from the anterior–posterior compartment, inferred using the manifold classifier trained on GFP-positive and GFP-negative cells. **b**, Top, anterior and posterior cells labelled by measured data (leftmost two columns). Anterior and posterior positions of AFP–GFP-positive and AFP–GFP-negative cells inferred (rightmost two columns) using probabilities in **a** (top panels). Bottom, GFP-positive and GFP-negative cells labelled by measured data (leftmost two panels). GFP-positive and GFP-negative status of the anterior–posterior compartment cells inferred (rightmost two panels) using probabilities in **a** (bottom panels). **c**, Top left, plot showing the first diffusion component of the E8.75 cells. Top middle, top right, plots showing the expression of anterior marker *Nkx2.1* and posterior marker *Hoxb9* in E8.75 cells. Bottom, bulk RNA-seq expression of *Nkx2-1* and *Hoxb9* in quadrants of the gut tube along the anterior–posterior axis compares with anterior–posterior single-cell expression patterns. **d**, Plot showing the proportion of anterior and posterior cells in bins along the anterior–posterior pseudo-space. **e**, Heat map showing Pearson's correlations between anterior–posterior pseudo-space orderings, determined using a varying number of diffusion components and highlighting the robustness of the ordering. **f**, Plots comparing the anterior–posterior pseudo-space ordering of GFP-positive and GFP-negative cells (replicate 2, 13,335 cells) generated de novo using only the replicate 2 cells (x axis, left) with the projected ordering from replicate 1 (8,143 cells) (y axis). Right, similar comparison

with the pseudo-space ordering determined using cells of both the replicates on the x axis. **g**, Same as **f**, for replicates of anterior–posterior cells (replicate 1, 1,821 cells; replicate 2, 1,691 cells). Plots show the Pearson's correlation. **h**, ROC for classification of E7.5 visceral and definitive endoderm cells (4,378 cells). **i**, Plots showing the expression patterns of genes that are best predictive of the definitive endoderm class in the visceral and definitive endoderm classifier (top, definitive endoderm; bottom, visceral endoderm). **j**, Plots showing the expression patterns of genes in the definitive endoderm that are best predictive of visceral endoderm class in the visceral and definitive endoderm classifier. **k**, Force-directed layouts following Harmony of E7.5 and E8.75 visceral and definitive endoderm cells, with E7.5 cells coloured in red (definitive endoderm) and blue (visceral endoderm) (left). E7.5 visceral and definitive endoderm cells coloured by the branch probability of anterior localization (middle) and posterior localization (right). Black arrowheads indicate early emergence of anterior–posterior spatial patterning at E7.5, with E7.5 definitive endoderm cells predominantly destined towards the anterior, and visceral endoderm cells predominantly destined towards the posterior. **l**, Three-dimensional renderings of gut tube, depicting all endoderm cells along the anterior–posterior axis. Nuclei of visceral and definitive endoderm cells are labelled in green and grey, respectively. **m**, Plots comparing the ranks of proportion of GFP-positive cells along anterior–posterior positioning in the AFP–GFP-embryo-derived *NeuroLucida* reconstructed gut tube replicates (x axis), and the ranks of visceral endoderm cell proportions in bins along the anterior–posterior pseudo-space axis (y axis); the anterior–posterior axis was partitioned into 20 bins; each dot represents the fraction of visceral endoderm cells in that bin.



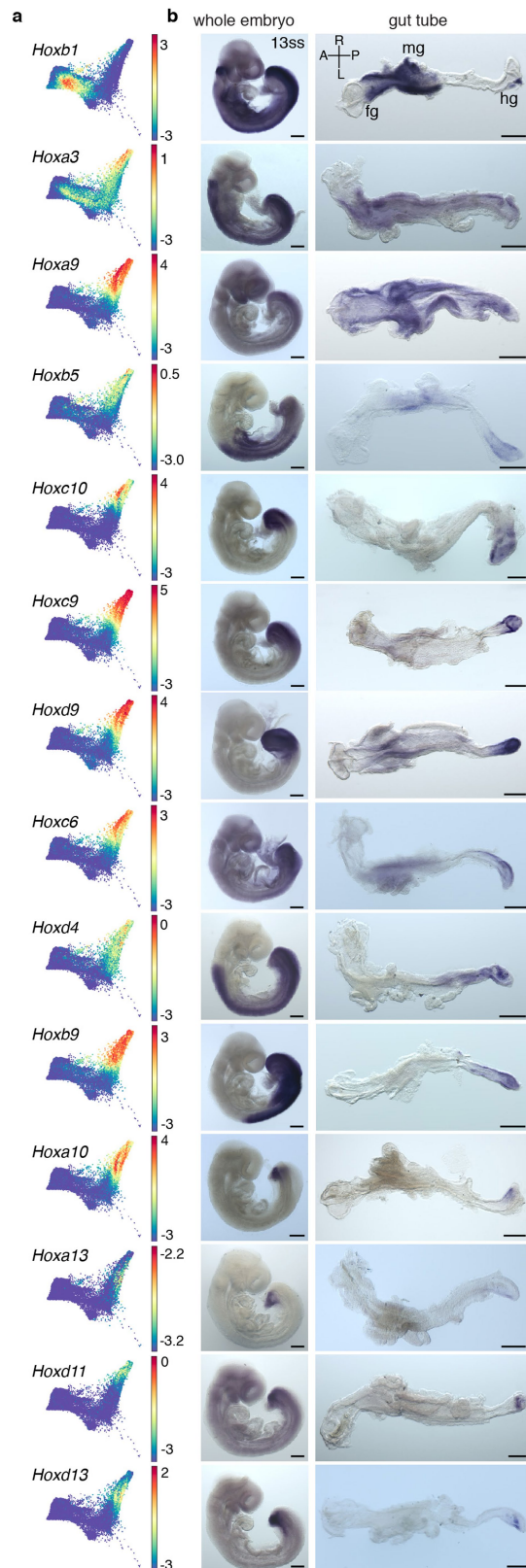
Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Spatial patterning of the gut tube at E8.75.

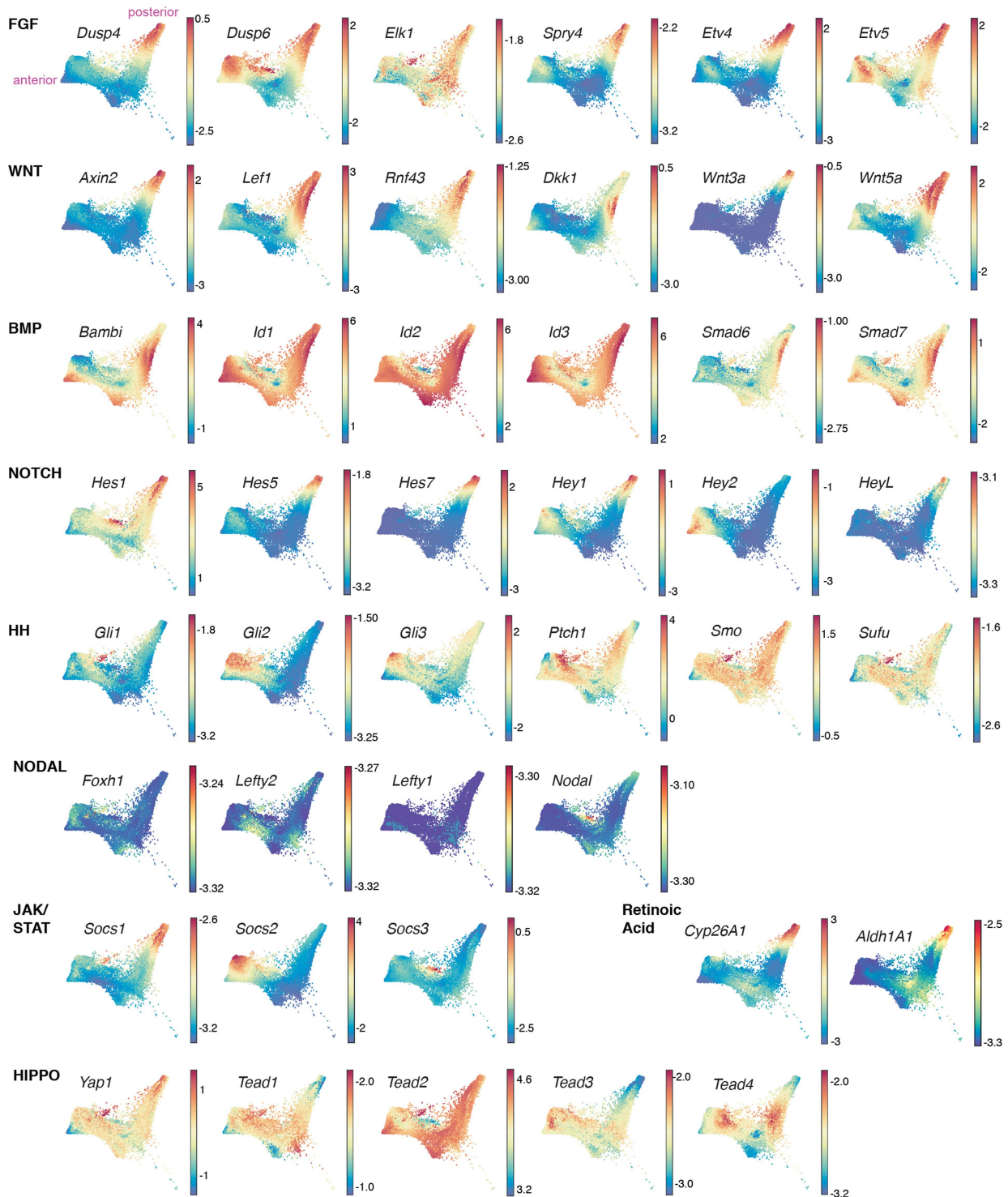
a, Plots showing individual Phenograph cluster densities of the E8.75 gut tube cells ordered along the anterior–posterior pseudo-space (left) and in force-directed layouts (middle). ISH of representative differentially expressed genes in each cluster on whole E8.75 embryos ($n > 3$ for each gene) or micro-dissected E8.75 gut tubes ($n > 3$ for each gene) (right). Arrowheads point to expression of representative gene for each particular cluster. Scale bars, 200 μm (except for *Nkx2-1*, 100 μm). no, notochord.

b, Density of E8.75 cells along the anterior–posterior pseudo-space axis. **c**, Comparison of empirical anterior–posterior pseudo-space axis

and the predicted anterior–posterior pseudo-space, using expression of transcription factors. Each dot represents the anterior–posterior pseudo-space of a cell computed by all genes, versus pseudo-space prediction by the selected transcription factors alone. **d**, Plot showing the ranking of different transcription factors according to their predictive power, on the basis of the regression model. **e**, Heat map showing the coefficients for the top transcription factors when different proportions of cells are subsampled for the regression (total cells, 24,990). **f**, Heat map showing the Pearson's correlation of transcription factor coefficients in **e**, highlighting the robustness of transcription factor coefficients in regression.



Extended Data Fig. 10 | Hox gene expression within the E8.75 gut tube.
a, Force-directed layouts showing Hox genes expressed in gut endoderm cells at E8.75. **b**, Whole-mount mRNA ISH on whole E8.75 embryos ($n > 3$ for each gene) and micro-dissected gut tubes ($n > 3$ for each gene) of Hox genes, depicting the distribution of Hox genes along the anterior–posterior axis. Scale bars, 100 μm (*Hoxc10* and *Hoxd11*), 200 μm (all other panels).



Extended Data Fig. 11 | Signalling map of the gut tube of the E8.75 mouse embryo. Force-directed layouts of context-independent targets of key signalling pathways acting within the endoderm lineage of the embryo.

Fibroblast growth factor (FGF); WNT; bone morphogenic protein (BMP); NOTCH; Hedgehog (HH); NODAL and TGF β signalling (NODAL); JAK and STAT; retinoic acid; and HIPPO.

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

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

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

ZEN (Carl Zeiss - <https://www.zeiss.com/microscopy/us/products/microscope-software/zen.html>), NeuroLucida (<https://www.mbfioscience.com/neuroLucida>); Adobe Illustrator and Photoshop: Version CC

Data analysis

Python 3.6, R 3.5, Harmony 0.1, Palantir 0.1, pandas 0.22, numpy 1.14.2, scipy 1.0.1, sklearn, fa2, matplotlib 2.2.2, seaborn 0.8.1, glmnet 2.0-16, DESeq2 1.22.1, networkx 2.1, Imapris (<http://www.bitplane.com/>)

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

All the generated scRNA-seq and bulk RNA-seq data will be deposited to GEO. Accession numbers: GSE123046 and GSE123124

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](https://www.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	No statistical methods were used to predetermine sample size. We used single cell RNA-seq in this study which by nature generates thousands of cell per sample. We therefore generated each sample in duplicate or triplicate and verified that the behavior of the cells from replicates is similar.
Data exclusions	No exclusion was applied.
Replication	We profiled 13 different samples, each collected in duplicate or triplicate, representing a total of 112,217 cells. We confirmed that all the replicates (except for E5.5 sample) were identical by examining the structure of the data (Supp. Fig. 2). E5.5 timepoints are extremely hard to synchronize and hence do not completely overlap
Randomization	Samples were collected from sequentially-staged wild-type embryos. No randomization was required since there were no clinical trials
Blinding	No blinding was required for our study since we did not perform any clinical trials

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Included in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

Methods

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Unique biological materials

Policy information about [availability of materials](#)

Obtaining unique materials All unique materials (e.g. mouse lines, Afp-GFP and Ttr-Cre) are available from the authors upon request.

Antibodies

Antibodies used	Details of all antibodies used in this study are provided in Materials and Methods section. Including, Chicken anti-GFP from Aves (cat# GFP 1020, lot# 511FP12), rabbit anti-RFP from Rockland (cat# 600-401-379, lot#39707), goat anti-hGata6 from R&D system (cat# AF1700, lot#KWT0316021), Rat anti-Sox2 (cat# eBioscience 14-9811-80, lot# 4347621, clone: Btjce), rabbit anti-Nanog from Cosmo Bio (cat#REC-RCAB002P, lot# C01QG10), Donkey anti-chicken Alexa Fluor® 488 from Jackson ImmunoResearch Laboratories (cat# 703-545-155), Donkey anti-rabbit Alexa Fluor® 488 from Invitrogen (cat# A21206, lot# 1834802), Donkey anti-goat Alexa Fluor® 647 from Invitrogen (cat# A21447, lot# 1841382), Donkey anti-rat DyLight® 650 from Invitrogen (cat# SA5-10029, lot# RL2316871), Donkey anti-rabbit Alexa Fluor® 546 from Invitrogen (cat# A10040, lot# 1833519), Donkey anti-goat Alexa Fluor® 568 from Invitrogen (cat# A11057, lot# 1871957) .
Validation	Antibodies were used on wild-type mouse embryos and validated on mouse null mutants (e.g. Gata6, see Schrode et al., Developmental Cell 2014) wherever possible.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	An H2B-tdTomato ES cell line described in Morgani et al., (Cell Reports 2013) was obtained from the laboratory of Josh Brickman (DanStem Institute, Copenhagen).
Authentication	No authentication procedure was used on this ES cell line.
Mycoplasma contamination	All ES cell lines used in this study have tested negative for mycoplasma.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used in this study.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Mice were maintained in accordance with the guidelines of the Memorial Sloan Kettering Cancer Center (MSKCC) Institutional Animal Care and Use Committee (IACUC) protocol no. 03-12-017 (to AKH).
Wild animals	The study did not use any 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	To obtain single cells from 13ss gut tubes, Afp-GFP/+/ embryos were dissected out of the uterus, and extra-embryonic membranes and heads were removed. Torsos washed in three drops of DMEM/F12 on ice and subjected to Pancreatin/Trypsin treatment (2.5% Pancreatin / 0.5% Trypsin in PBS) for 5 min (exact time was batch dependent and empirically tested) on ice and then washed in three drops of DMEM/F12, 10% Newborn calf serum on ice. Gut tubes were dissected out using Tungsten needles (FST Cat No. 10130-10) and washed in cold DMEM/F12. Gut tubes were then incubated for 20 min at 37°C in Accutase/0.25% Trypsin (1:2) for the dissociation of single cells. To obtain single cells from definitive and visceral endoderm cell from stages E7.5 and visceral endoderm cells from E6.5, embryos were washed in three drops of DMEM/F12 on ice and subjected to Pancreatin/Trypsin treatment (2.5% Pancreatin / 0.25% Trypsin in PBS) for 3 min (E7.5) and 45s (E6.5) on ice and then washed in three drops of DMEM/F12, 10% Newborn calf serum on ice. The endoderm layer was removed from the rest of the tissue using Tungsten needles and subsequently washed in cold DMEM/F12 and incubated for 20 min at 37°C in Accutase/0.25% Trypsin (1:2). For E5.5 (defined as the stage when the DVE/AVE was distally positioned, observed as a thickening of the emVE epithelium), whole embryos were collected, and Reichert's membrane removed with Tungsten needles. Embryos were washed in cold DMEM/F12 and then incubated in 0.25% Trypsin for 5 min at 37°C. To dissociate the tissue into a single cell suspension, DMEM/F12, 20% Newborn calf serum, 4mM EDTA were added in 1:1 ratio. Cell clumps were dissociated into single cells using a mouth pipet and 75mm glass capillaries. Subsequently, the single cell suspension was filtered through a FlowMI cell strainer (40µm) to remove debris. Single cells were spun down at 450g for 4 min at room temperature and then cell numbers were determined using a Neubauer hemocytometer.
Instrument	FACSAria IIu SORP (BD Biosciences) with a 488 nm, 561 nm, and 405 nm laser to excite GFP, ethidium homodimer-1, and Calcein Violet, respectively, and using a 530/50, 582/15, and 450/50 nm band pass filters to detect these same fluorochromes, respectively.
Software	Data was acquired and sorted using FACSDiva software (ver. 8.01, BD Biosciences), and analysis was done using FCS Express 6 (ver. 6.06, De Novo Software)
Cell population abundance	From each tissue we were able to sort population a few thousand cells. When possible, purity checks were performed, by sorting a thousand cells and acquiring up to 400 events. Purity was above 99%.
Gating strategy	In all cases the SSC vs FSC plot showed a clear population of cells and a gate was drawn to exclude small debris. SSC-H vs SSC-W as well as FSC-H vs FSC-W plots were used to exclude aggregates. Endoderm and gut tube single cell suspensions were sorted based on GFP content, with both GFP positive and negative fractions being collected, excluding dead cells using ethidium

homodimer-1. A negative control (WT embryos) was used to define GFP positive events. Yolk sac and ParE single cell suspensions were sorted excluding debris by selecting Calcein Violet positive events and excluding dead cells with ethidium homodimer-1.

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