



ARTICLE

Single-cell patterning and axis characterization in the murine and human definitive endoderm

Lin-Chen Li^{1,2,3}, Xin Wang¹, Zi-Ran Xu^{1,4}, Yan-Chun Wang⁵, Ye Feng^{1,4}, Liu Yang^{1,2}, Wei-Lin Qiu^{1,4}, Li Yang^{1,2}, Xin-Xin Yu¹, Jun Gu³ and Cheng-Ran Xu¹

Defining the precise regionalization of specified definitive endoderm progenitors is critical for understanding the mechanisms underlying the generation and regeneration of respiratory and digestive organs, yet the patterning of endoderm progenitors remains unresolved, particularly in humans. We performed single-cell RNA sequencing on endoderm cells during the early somitogenesis stages in mice and humans. We developed molecular criteria to define four major endoderm regions (foregut, lip of anterior intestinal portal, midgut, and hindgut) and their developmental pathways. We identified the cell subpopulations in each region and their spatial distributions and characterized key molecular features along the body axes. Dorsal and ventral pancreatic progenitors appear to originate from the midgut population and follow distinct pathways to develop into an identical cell type. Finally, we described the generally conserved endoderm patterning in humans and clear differences in dorsal cell distribution between species. Our study comprehensively defines single-cell endoderm patterning and provides novel insights into the spatiotemporal process that drives establishment of early endoderm domains.

Cell Research (2020) 0:1–19; <https://doi.org/10.1038/s41422-020-00426-0>

INTRODUCTION

Following the gastrulation stage of animal development, the embryo establishes the basic axes of the body, and individual organs begin to develop within the defined regions of the newly formed germ layers known as the ectoderm, mesoderm, and definitive endoderm (DE).¹ DE contributes to the respiratory system and gastrointestinal tract, including their associated organs, such as the liver and pancreas.^{2,3} Understanding the spatial and temporal landscapes of DE formation is critical for uncovering the regulatory mechanisms underlying cell lineage specification of the earliest progenitors in endoderm-derived organs. However, the precise spatial and temporal endoderm patterning remains undefined.

In the mouse embryo, DE formation begins during gastrulation. Around embryonic day (E) 7.25, the embryo assumes a cup shape, and a single-cell sheet of DE is located at the outermost surface that lines the mesoderm.⁴ Recent studies have revealed the molecular genealogy of the germ layers.^{5,6} By E8.0, pockets develop at the anterior and posterior ends of this epithelial sheet to form an anterior intestinal portal (AIP) and a caudal intestinal portal (CIP), respectively. The leading edge of the ventral intestinal pocket forms a “lip” structure. Then, with the extension of AIP and CIP, the AIP lip (AL) and the CIP lip merge to form the gut tube. During these processes, the anterior–posterior (A–P), dorsal–ventral (D–V), and medial–lateral (M–L) axes are established.⁷ The gut tube becomes regionalized and is typically divided into the foregut (FG), midgut (MG), and hindgut (HG). The

FG region is generally considered to give rise to the esophagus, stomach, thyroid, trachea, lungs, liver/biliary system, and pancreas; the MG produces the small intestine, and the HG develops into the colon.^{2,8} However, the molecular criteria and boundaries that define these critical regions have not been determined.

The endoderm sheet has begun to regionalize in the early somitogenesis stage (~E8.25).⁹ In this stage, endoderm cells are in direct contact with mesoderm and neurectoderm tissues, which secrete patterning signals, including fibroblast growth factors (FGFs),^{10,11} Wnt,¹² bone morphogenetic proteins (BMPs),^{13–15} and retinoic acid (RA).¹⁶ Together, these signals promote and maintain posterior endoderm fate but repress anterior identity.^{11,12,17} A limited number of transcription factors (TFs) define and regulate regionalization along the A–P axis. *Sox2* and *Cdx1/2/4* are specifically expressed in the anterior and posterior regions of the endoderm, respectively.² FG identity is regulated by *Sox2*, *Hhex*, *Nkx2.1*, and *Foxa2*, whereas dose-dependent activity of *Cdx* genes determines cell fate in the MG and HG as well as the position of the boundary that divides the FG and the MG/HG.^{2,18–21} The cells that comprise the regionalized endoderm are segregated into a number of primitive organ subdomains, which illustrates the heterogeneous developmental potential of endoderm cells. However, this endoderm patterning has not been clearly defined.

D–V patterning is critical for separation of the dorsal organs, such as the esophagus and dorsal pancreas, from the ventral organs, such as the liver, thyroid, thymus, and ventral pancreas, in

¹Ministry of Education Key Laboratory of Cell Proliferation and Differentiation, College of Life Sciences; Department of Human Anatomy, Histology, and Embryology, School of Basic Medical Sciences; Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China; ²Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China; ³Beijing Institute of Collaborative Innovation, Beijing 100094, China; ⁴PKU-Tsinghua-NIBS Graduate Program, Peking University, Beijing 100871, China and

⁵Haidian Maternal & Child Health Hospital, Beijing 100080, China

Correspondence: Cheng-Ran Xu (cxu@pku.edu.cn)

These authors contributed equally: Lin-Chen Li, Xin Wang, Zi-Ran Xu, Yan-Chun Wang

Received: 7 May 2020 Accepted: 28 September 2020

Published online: 26 October 2020

the corresponding DE regions. Notably, the pancreas is separately derived from both a dorsal and ventral endoderm domain, and these two domains receive distinct inductive signals from adjacent tissues and follow separate pathways to differentiate into pancreatic progenitors (PPs).^{22–25} However, the programs that direct generation of these two pancreatic domains are poorly understood. In addition, the processes that guide D–V specification and patterning have not been well studied.

In addition to the pancreas, the liver originates from separate endodermal domains. On the AL, the cells in the paired lateral domains develop only into the liver, while the cells in the midline domain generate the liver, thyroid, and pancreas.^{26,27} These findings indicate that a fraction of endoderm cells specify along the M–L axis. Such patterning of liver and pancreas has also been observed during zebrafish development.²⁸ However, little is known about the regulation of M–L axis formation.

Several protocols have been developed to induce differentiation of human pluripotent stem cells (hPSCs) into lineage-biased endoderm cells, which can then be further differentiated into organ-specific progenitors.^{1,29} However, our understanding of human endoderm development is very limited.

Single-cell RNA sequencing (scRNA-seq) is a powerful tool that enables global observations of cell lineage composition and developmental pathways. scRNA-seq is often performed using a droplet-based high-throughput approach, such as the 10× Genomics Chromium platform, or the well-based Smart-seq2 method,³⁰ in which cDNA libraries are produced from single cells deposited into individual wells on a plate. The 10× Genomics system allows for a high-throughput approach with high cell coverage and was used to map the transcriptional landscape of early organogenesis in mice.^{31,32} However, due to technological limitations of the 10× Genomics platform, mainly its higher noise and lower sensitivity for low-abundance transcripts,^{33,34} early specified lineages are difficult to be identified, which affects subsequent analyses of the patterning and developmental pathways of endoderm lineages. To avoid these limitations, we utilized a more sensitive Smart-seq2 and a modified STRT-seq (mSTRT-seq) protocols^{30,35} to investigate microdissected endoderm cells during early somitogenesis in mice and humans, and defined the single-cell patterning and axis characterization of the cell subpopulations in the DE.

RESULTS

Four major regions of the DE

To characterize the composition and spatial distribution of the cells that comprise the epithelial sheet of the endoderm, we performed scRNA-seq on endoderm cells purified by fluorescence-activated cell sorting (FACS) at the 9-somite stage (SS) (E8.5). At this stage, patterning of the endoderm sheet is almost complete, but most endoderm-derived organogenesis processes have not yet begun. We dissociated the whole gut tube by trypsinization and microdissected the tissue as previously described,³⁶ and carefully removed the remaining adhesive extraembryonic and mesoderm tissues (Fig. 1a). To retain information regarding the spatial relationship of the cells in the endoderm, we also carefully dissected various parts of the endoderm tissue from whole embryos, including ① the ventral AIP, ② the dorsal anterior AIP (from the foremost end to the position before the first somite), ③ the dorsal region at the 1st–2nd somite positions, ④ the medial (M), left (L), and right (R) lateral regions (④M, ④L, and ④R) of the AL (Supplementary information, Fig. S1a), ⑤ the medial and lateral regions (⑤M, ⑤L, and ⑤R) of the endoderm at the 3rd–9th somite positions (Supplementary information, Fig. S1b), and ⑥ the rest of the posterior endoderm (Fig. 1b, c). The endoderm cells from the dissected tissues were enriched by FACS using an antibody specific for the epithelial cell marker EpCAM.³⁷ To achieve high transcript coverage in individual cells, we used a well-based scRNA-seq method, Smart-seq2, to obtain the full-length

transcriptomes.³⁰ A total of 1772 single-cell transcriptomes passed the quality control criteria. We obtained >0.3 million mapped reads per cell, and on average, detected 9000 genes in each cell (Supplementary information, Fig. S1c, d, Table S1).

Cell types were identified by marker gene expressions. After excluding cells of mesoderm (*Foxf1*⁺*Hand1*⁺), endothelial (*Sox7*⁺*Icam2*⁺), notochord (*Nog*⁺*Lmx1a*⁺), paraxial mesoderm (*Mesg1*⁺*Fgf4*⁺), yolk sac endoderm (*Cubn*⁺*Apoc2*⁺), primordial germ cell (*Dppa3*⁺*Pou5f1*⁺), and ectoderm (*Tfap2a*⁺*Gjb3*⁺*Dlx5*⁺), the remaining 1314 cells, expressing the endoderm epithelium markers *Epcam*, *Foxa1*, *Foxa2*, and *Cxcr4*, were classified as endoderm cells (Supplementary information, Fig. S1e, f). Samples from individual embryos were pooled, and scRNA-seq analyses of samples dissected from different parts of the endoderm tissue were independently repeated at least twice. T-distributed stochastic neighbor embedding (t-SNE) plots did not reveal obvious batch effects between biological replicates of intact whole endoderm samples, and the cells from different portions of the endoderm tissue were distributed within the range of cells from the whole endoderm (Supplementary information, Fig. S1g).

To identify the endoderm cell populations, single cells were classified using gene co-expression network-based clustering (GCN clustering) analysis,^{38,39} an algorithm based on the assumption that cells with the same identity should display strong correlations in their gene expression profiles and thus co-express the same set of genes. GCN clustering analysis identified four cell types (Supplementary information, Fig. S1h, Table S2), which specifically expressed distinct clusters of GCN-associated genes, as shown in a heatmap (Supplementary information, Fig. S1i) and a GCN plot (Fig. 1d). The distribution patterns of the four cell types are displayed on a t-SNE plot (Fig. 1e). A transcriptomic level differential expression analysis revealed four cell type-specific gene groups (A–D) (Fig. 1f; Supplementary information, Table S2). To investigate the spatial patterns of gene expression in the endoderm germ layer, we performed whole-mount in situ hybridization (ISH) on 9-SS embryos using antisense riboprobes against the group-specific genes. The group-A genes *Pax9* and *Has2* were expressed in the AIP region, except for the lip area (Fig. 1g), and cells dissected from portions ① and ② were enriched in cluster I (Fig. 1e). Therefore, we concluded that the cluster-I cells were FG cells, which span the endoderm from its most anterior portion to the position before the first somite dorsally and to the position before the AL ventrally (Fig. 1h). A group-B gene, *Hhex*, was expressed in the AL region, and most cells dissected from the ④M, ④L, and ④R regions were included in cluster II (Fig. 1e, g), which confirmed that the cluster-II cells were AL cells. A group-C gene, *Nepn*, was detected in the MG region³⁶ (Fig. 1g), and the cells in regions ③ and ⑤ (M, L, and R regions) were included in the cluster-III cells (Fig. 1e). Thus, cluster-III cells, which dorsally spanned from the position of the first somite to the position of the last somite, were identified as MG cells. Similar analyses showed that the group-D genes *Cdx2* and *Wnt5b* were expressed in the region posterior to the last somite, and most of the region ⑥ cells were included in cluster IV (Fig. 1e, g). Therefore, cluster-IV cells were designated as HG cells. Together, single-cell transcriptomic analyses combined with microdissection and ISH delineated four major cell populations and their spatial patterns on the 9-SS endoderm (Fig. 1h). We next focused on each major cell population to identify subpopulations and their distribution patterns in the DE.

To eliminate loss of cells during well-based scRNA-seq, we dissected endoderm tissue from 9-SS mouse embryos and conducted unbiased scRNA-seq using the 10× Genomics platform. Generally, ~1500–2000 genes were detected in each cell (Supplementary information, Fig. S2a–c, Table S1). After filtering poor-quality cells, 11,438 cells were retained for further analyses. After removing other cell types based on their marker gene expressions, we obtained 6498 endoderm cells (Supplementary information, Fig. S2d, e). We used the mutual nearest neighbor (MNN) algorithm⁴⁰ to correct for methodological batch effects and

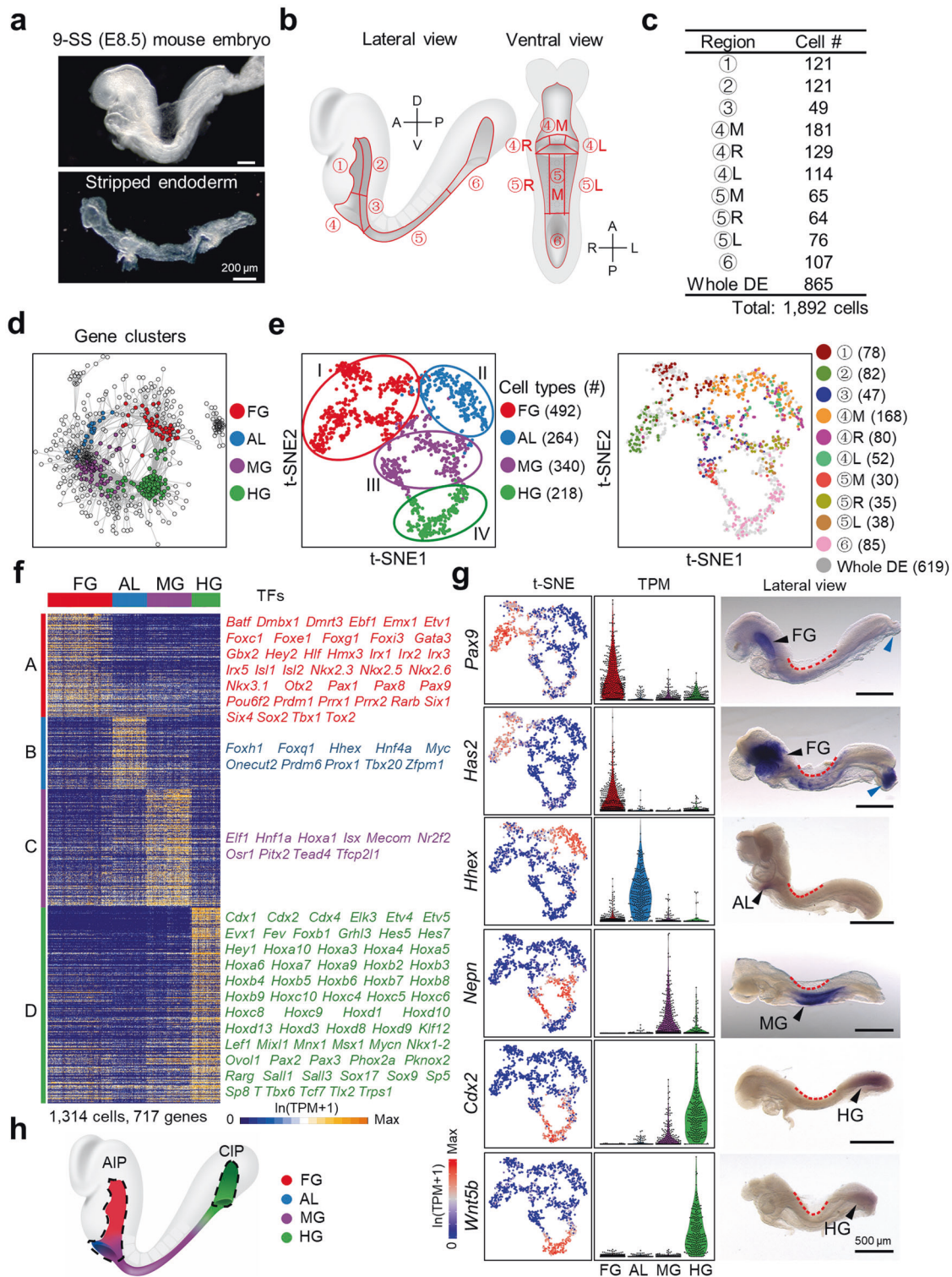


Fig. 1 scRNA-seq identified four distinct major populations of endoderm cells. **a** The endoderm gut tube (lower) was dissected from an E8.5 (9-SS) mouse embryo (upper). Scale bars, 200 μ m. **b** Schematic representation of dissected endoderm regions. R, right; M, medial; L, left. **c** Overview of the single-cell number (#) from specific endoderm regions. The circled numbers represent the different regions indicated in **b**. **d** Distinct gene clusters differentially expressed in endoderm cells. Each dot represents a GCN-associated gene. The cell type-specific gene clusters are colored. **e** The t-SNE plot shows four distinct cell types (left) or cell sources (right). Each dot represents a single cell. Cell counts are indicated in brackets. **f** Heatmap of 717 DEGs identified in four distinct endoderm cell populations and four groups (A–D) of genes. Each column represents a single cell, and each row represents one gene. Group-specific TFs are listed on the right. **g** Expression levels of marker genes are shown on the t-SNE plots (left) and violin plots (TPM: transcripts per million), middle). A dot within the t-SNE plot or violin plot represents a single cell. Validation of marker genes (right) by whole-mount ISH on 9-SS embryos (signal indicated by an arrowhead). Blue arrowhead indicates the allantois region. Each segment of the dashed line represents a somite pair. $n > 3$ for each gene. Scale bars, 500 μ m. **h** Schematic representation of the spatial distribution of endoderm cell populations. Dotted lines indicate the AIP and CIP.

projected the previous Smart-seq2-analyzed cells onto the 10× Genomics t-SNE plot of 9-SS cells, as well as 13-SS cells generated by others.³¹ We found that Smart-seq2 cells were distributed throughout the 10× Genomics plot (Supplementary information, Fig. S2f), indicating that no cell population was omitted by our well-based scRNA-seq. However, the 10× Genomics platform using both the standard analytic pipeline and the pseudocell method⁴¹ identified a much smaller number of differentially expressed genes (DEGs) than Smart-seq2, although most of the genes identified by 10× Genomics overlapped with those identified by Smart-seq2 (Fig. 1f; Supplementary information, Fig. S2g–l).

Patterning of five cell subpopulations in the FG

Using GCN clustering, we identified five gene clusters, which represented five subpopulations of FG cells (FG.1–5) (Fig. 2a, b; Supplementary information, Table S2). Differential expression analysis classified 9 differential gene groups in FG.1–5 cells (Fig. 2c; Supplementary information, Table S2). Genes in groups A–E were generally specific to a single subpopulation, whereas genes in groups F–I were expressed in at least two subpopulations. To verify the stability of our FG cell classifications, we performed a random sampling analysis of FG cells and found that the GCN was stable and FG cells remained divided into five subpopulations even when the number of cells was reduced to 100 (Supplementary information, Fig. S3a, b).

We next examined the spatial expression patterns of these subpopulation-specific genes using ISH to determine the distributions of FG.1–5 cells on the FG endoderm. *Smoc2* and *Foxg1* were specifically expressed in FG.1 cells, and we observed *Smoc2*⁺ and *Foxg1*⁺ cells in the most rostral region of the FG, forming a cap-like shape with an arched bulge at the top toward the head fold (Fig. 2d). FG.2 cells, which specifically expressed *Cdc42ep3*, were located in the region between FG.1 and the middle position of the FG along the A–P axis. Within this segment, FG.2 cells covered the entire dorsal and lateral ventral regions but not the medial ventral region (Fig. 2d). ISH against *Prrx2*, an FG.3-specific gene, revealed that FG.3 cells were located in the rostral portion of the FG. Similar to FG.2 cells, FG.3 cells were present throughout the dorsal and lateral ventral regions but not in the medial ventral region (Fig. 2d). The FG.4-specific gene *Pyy* was expressed in the cells located along the ventral midline adjacent to the AL. *Nkx2.1* was exclusively expressed in FG.5 cells, and *Nkx2.1*⁺ cells were located in the medial ventral region of the rostral part of the FG (Fig. 2d). Using genes expressed in at least two subpopulations, we determined the relative spatial relationships between the subpopulations. We performed ISH against *Cxcl12* (FG.1 and FG.2), *Fgf8* (FG.2 and FG.3), *Tbx1* (FG.1 and FG.3), and *Nkx2.3* (FG.2, FG.4, and FG.5) and confirmed the distributions of each subpopulation (Fig. 2d). Interestingly, immunofluorescence analysis of the endoderm marker FOXA2 revealed that at the 9-SS, the dorsal FG forms a single cell layer with a squamous shape,⁴² the lateral cells become columnar, causing the layer to thicken, and the cells at the ventral midline become stratified or pseudostratified (Supplementary information, Fig. S3c). This phenomenon also explained why ISH signals were easier to be observed in the lateral or ventral FG than in the dorsal FG. Consistent with the patterning of FG.1–5 (Fig. 2e), we found that the cells dissected from the dorsal portion of the AIP (②, Figs. 1b and 2b) contributed solely to populations FG.1–3, whereas cells from the ventral portion of the AIP (①, Figs. 1b and 2b) were mainly distributed in populations FG.4–5, with a small portion in populations FG.1–3. Therefore, at the 9-SS, FG cells are specified into five cell subtypes along three-dimensional (3D) axes.

Stratified patterning of the AL endoderm

GCN clustering analysis identified three subpopulations of AL cells (AL.1–3), which expressed distinct gene clusters (Fig. 3a; Supplementary information, Table S2). These cell clusters were

displayed on the t-SNE plot (Fig. 3b). We classified three groups of genes that were heterogeneously expressed in AL.1–3 cells (Fig. 3c; Supplementary information, Table S2). Notably, group-A and group-B genes, which were highly expressed in AL.1 and AL.3 cells, respectively, were also expressed in AL.2 cells, but at relatively lower levels; no AL.2-specific genes were identified. This finding suggested that AL.2 cells were in an intermediate state between AL.1 and AL.3 cells. To explore the spatial distribution of AL.1–3 cells, we performed ISH on 9-SS whole embryos using probes against group-specific genes, followed by sagittal sectioning. *Hhex* was expressed in all AL subpopulations (Figs. 1g and 3d). On the sagittal sections, we clearly observed that *Hhex*⁺ cells were distributed starting from the AIP opening and covered the posterior edge of the atrium and ventricle of the developing heart (Fig. 3d). *Cldn11* and *Upk3a* were specifically expressed in AL.1, and ISH showed that AL.1 cells were distributed linearly along the left–right axis (Fig. 3e, f). On the sagittal sections, AL.1 cells were adjacent to the narrow AIP opening (Fig. 3e). Similarly, by examining the expression patterns of the AL.3-specific genes *Ly6h*, *Dab2*, and *Ghrl*, we found that the linearly distributed AL.3 population was the outermost located population within the AL (Fig. 3e, f). *Afp* expression was lower in AL.2 cells than that in AL.3 cells, and we observed that AL.2 cells with lower *Afp* expression were located between AL.1 and AL.3 cells (Fig. 3e). Altogether, the distribution of AL cells form a fan-shaped structure that gradually widens from the innermost AL.1 to the outermost AL.3 stripe (Fig. 3g).

To further verify this stratified distribution of cells in the AL, we analyzed scRNA-seq data of cells dissected from ④M, ④L, and ④R (Fig. 1b). On the t-SNE plot, cells from portion ④M were distributed in the AL.1–3 populations, whereas those from ④L and ④R were generally associated with the AL.2 and AL.3 populations (Fig. 3b). This result was consistent with the observed fan-shaped stratification of AL.1–3 cells. Notably, ④L and ④R cells were intermingled on the t-SNE plot, suggesting that at the 9-SS, cells from the left and right portions of the AL were homogenous at the transcriptomic level, although these two cell groups will later contribute to different regions of the gut tube.⁴³ Together, our analyses identified three AL subpopulations and their stratified distribution patterns.

Segmentation and medial specification of the MG and HG

Similarly, using GCN clustering, we divided MG cells into three subpopulations (MG.1–3), which specifically expressed three groups of genes (Fig. 4a–c; Supplementary information, Table S2). The group-A gene *Hoxb1* was MG.1 specific and was expressed in the lateral region of the MG at the level of the first two somite positions (Fig. 4d). Notably, FG.3 cells, which were anteriorly adjacent to MG.1, also expressed *Hoxb1* at a lower level (Fig. 4d). *Upk1a* was specifically expressed in MG.2 cells, and whole-mount ISH and cross-sections showed that *Upk1a*⁺ cells were located in the lateral region in the segment between MG.1 and the position of the last somite (Fig. 4d). We also detected an MG.3-specific gene, *Mnx1*, that was expressed in the cells located on the MG midline (Fig. 4d). Consistent with the patterning of the MG subpopulations, we found that cells dissected from the dorsal region underneath the first two somites (③, Fig. 1b) contributed to the MG.1 and MG.3 populations (Figs. 1b and 4b) but not MG.2; the cells dissected from ⑤M were solely located in the MG.3 cluster (Figs. 1b and 4b), whereas cells from ⑤L and ⑤R belonged to the MG.2 subpopulation and did not display left–right heterogeneity at the transcriptomic level (Fig. 4b).

HG cells were further divided into two subpopulations (HG.1–2) (Fig. 4e–g; Supplementary information, Table S2). *Cxcl12* was highly expressed in HG.1 cells and its expression was detected in the lateral and ventral regions of the CIP, but not the tip region of the CIP (Fig. 4d). *Mnx1*, which was expressed on the midline of the MG (Fig. 4c, d), was highly expressed in HG.2 cells, indicating a

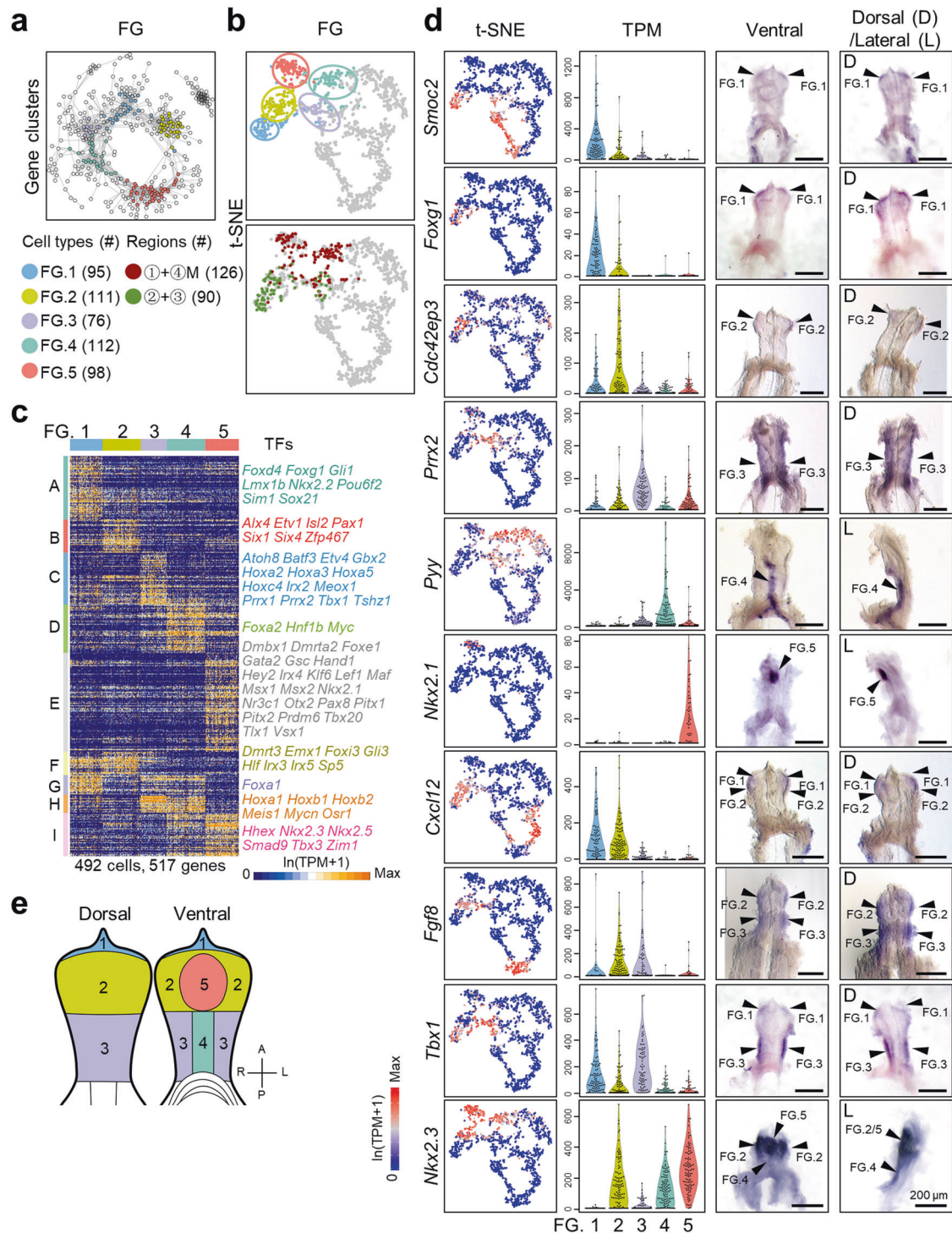


Fig. 2 Identification of FG subpopulations. **a** Distinct gene clusters differentially expressed in FG cells. Each dot represents a GCN-associated gene. The cell type-specific gene clusters are colored. **b** The t-SNE plot shows five distinct cell types (upper) or regional information (lower). Each dot represents a single cell. Cell counts are indicated in brackets. **c** Heatmap of 517 DEGs identified five distinct FG cell types and nine groups (A–I) of genes. Each column represents a single cell, and each row represents one gene. Group-specific TFs are listed on the right. **d** Expression levels of marker genes are shown on the t-SNE plots (left) and violin plots (TPM, middle). A dot within the t-SNE plot or violin plot represents a single cell. Validation of marker genes (right) by whole-mount ISH on 9-SS embryos (signal indicated by an arrowhead). $n > 3$ for each gene. Scale bars, 200 μm . **e** Schematic representation of the spatial distribution of FG cell types.

contiguous relationship between MG.3 and HG.2 cells. Whole-mount ISH against *Mnx1*, followed by cross sectioning, showed that HG.2 cells were indeed located in the medial of the HG and extended to the entire CIP tip (Fig. 4d, g). A similar pattern was

observed for another HG.2 gene, *Fgf8* (Fig. 4d). Together, these analyses identified the main segments of the MG and HG along the A–P axis and the specified midline region that extends through the MG and dorsal HG (Fig. 4h).

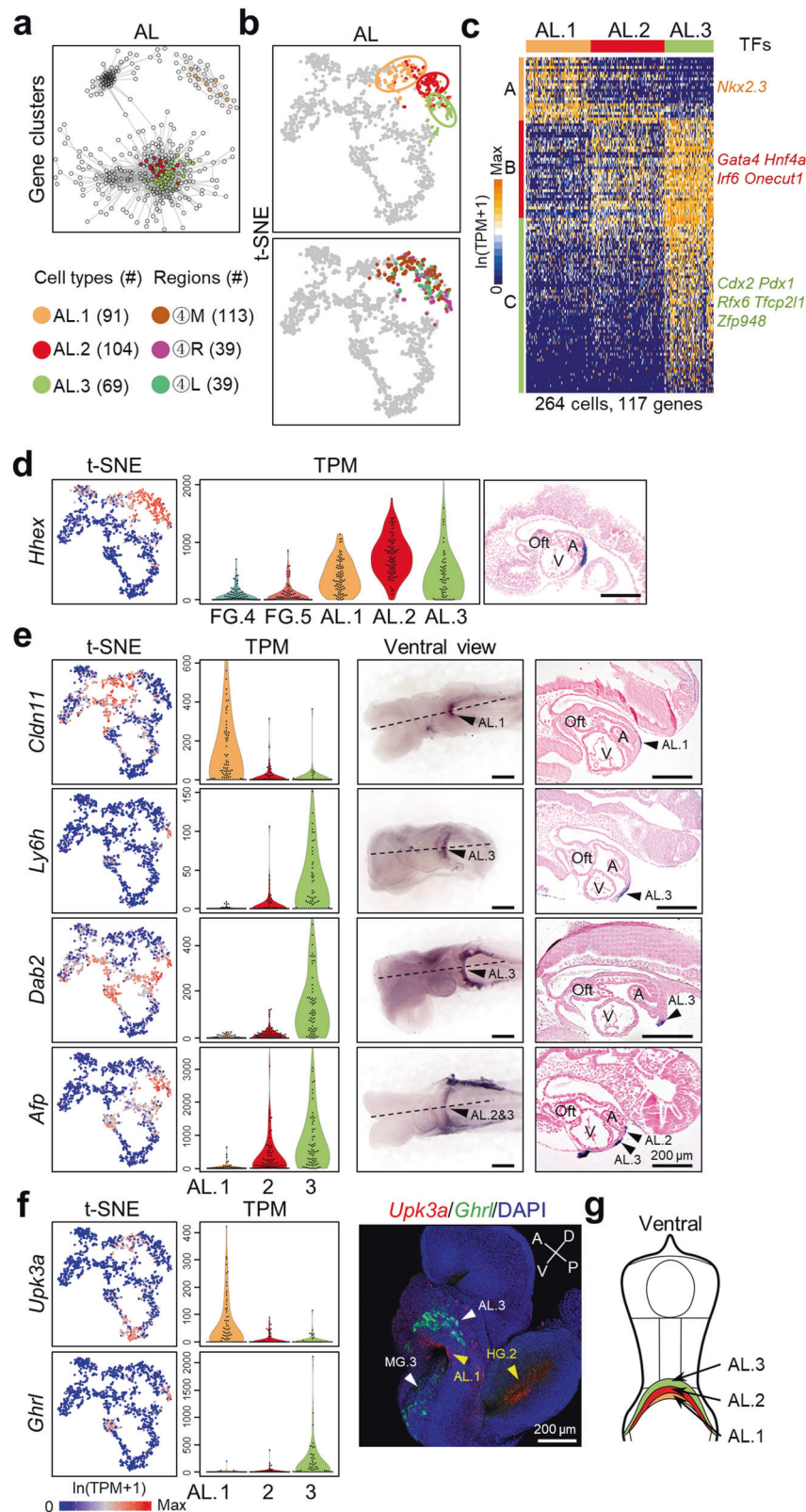
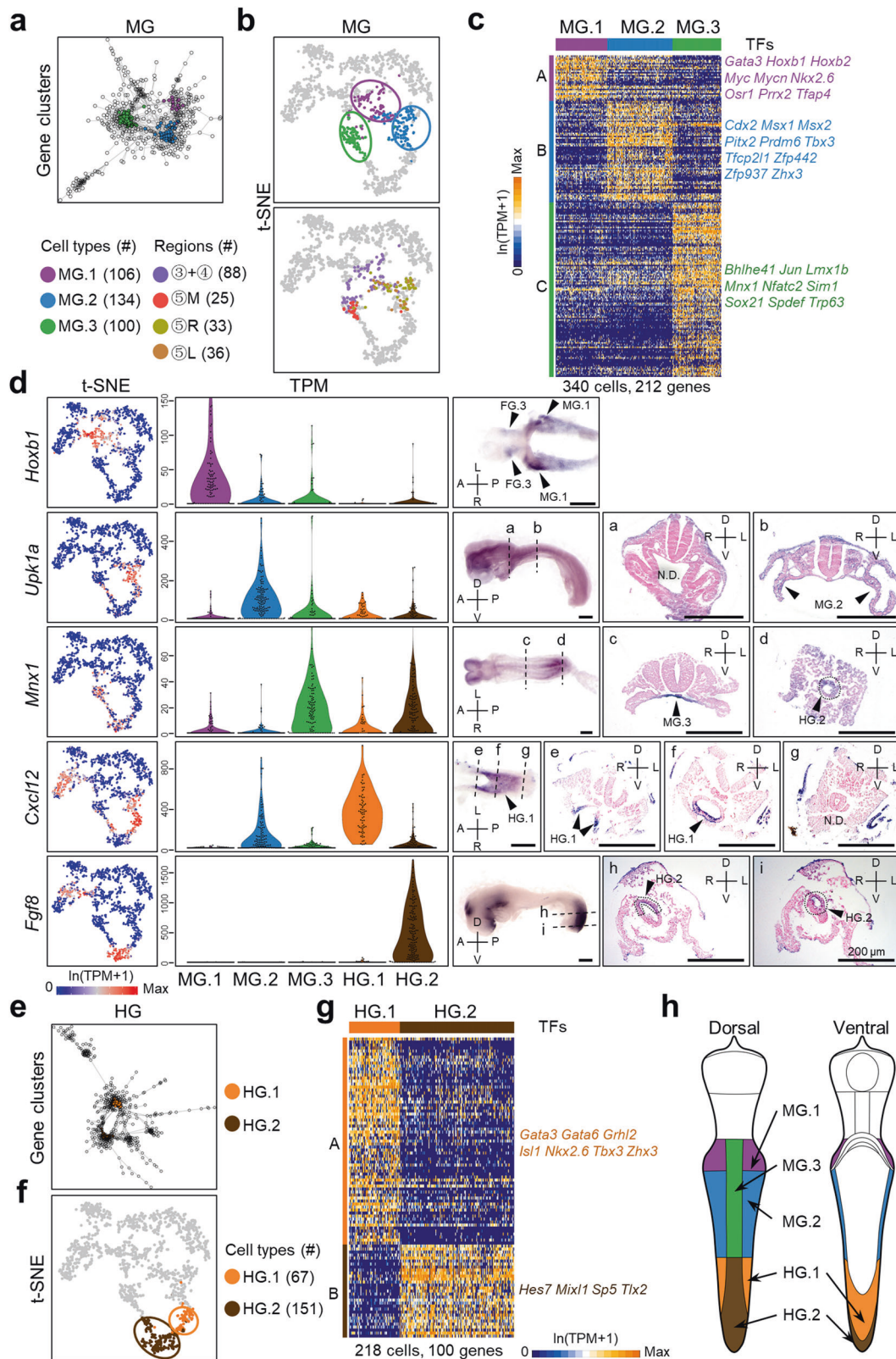


Fig. 3 Identification of AL cell types. **a** Distinct gene clusters differentially expressed in AL cells. Each dot represents a GCN-associated gene. The cell type-specific gene clusters are colored. **b** The t-SNE plot shows three distinct cell types (upper) or regional information (lower). Each dot represents a single cell. Cell counts are indicated in brackets. **c** Heatmap of 117 DEGs identified three distinct AL cell types and three groups (A–C) of genes. Each column represents a single cell, and each row represents one gene. Group-specific TFs are listed on the right. **d–f** Expression levels of marker genes are shown on the t-SNE plots (left) and violin plots (TPM, middle). A dot within the t-SNE plot or violin plot represents a single cell. Validation of marker genes (right) by whole-mount ISH (**d, e**) or single-molecule FISH (**f**, white arrowhead, *Ghrl*; yellow arrowhead, *Upk3a*). The dotted line indicates the site of histological sections. The signal is indicated by an arrowhead. Off, outflow tract; V, ventricle; A, Atrium. $n > 3$ for each gene. Scale bars, 200 μ m. **g** Schematic representation of the spatial distribution of AL cell types.



Patterning along the A-P, D-V, and M-L axes
Using scRNA-seq and ISH, we have clearly identified the spatial distribution of the earliest specified endoderm cell subpopulations (Fig. 5a). To uncover the molecular associations between these populations, we performed 3D force-directed layout (FDL) and

uniform manifold approximation and projection (UMAP) analyses and used the principal curve⁴⁴ to depict the ordering of cell populations, which reflects their spatial distributions on the endoderm layer (Fig. 5b, c; Supplementary information, Fig. S4a). Combined with the endoderm cell distribution map (Fig. 5a), we

Fig. 4 Identification of MG and HG cell types. **a** Distinct gene clusters differentially expressed in MG cells. Each dot represents a GCN-associated gene. The cell type-specific gene clusters are colored. **b** The t-SNE plot shows three distinct cell types (upper) or regional information (lower). Each dot represents a single cell. Cell counts are indicated in brackets. **c** Heatmap of 212 DEGs identified three distinct MG cell types and three groups (A–C) of genes. Each column represents a single cell, and each row represents one gene. Group-specific TFs are listed on the right. **d** Expression levels of marker genes are shown on the t-SNE plots (left) and violin plots (TPM, middle). A dot within the t-SNE plot or violin plot represents a single cell. Validation of marker genes by whole-mount ISH (right). The dotted line indicates the site of histological sections. The signal is indicated by an arrowhead and a dotted circle. N.D., no signal detected. $n > 3$ for each gene. Scale bars, 200 μm . **e** Distinct gene clusters differentially expressed in HG cells. Each dot represents a GCN-associated gene. The cell type-specific gene clusters are colored. **f** The t-SNE plot shows two distinct cell types. Each dot represents a single cell. Cell counts are indicated in brackets. **g** Heatmap of 100 DEGs identified two distinct HG cell types and two groups (A–B) of genes. Each column represents a single cell, and each row represents one gene. Group-specific TFs are listed on the right. **h** Schematic representation of the spatial distribution of MG and HG cell types.

revealed three paths of cell ordering along the A–P axis (Fig. 5b), which include the cell populations FG.2–FG.3–MG.1–MG.2–HG.1 (line-I), FG.5–FG.4–AL.1–AL.2–AL.3 (line-II), and FG.1–MG.3–HG.2 (line-III) (Fig. 5b). Although FG.1 was not directly connected to MG.3 on the FDL plot, when we compared FG.1-featured genes to MG.1–3-featured genes (Figs. 2c and 4c), we found that 22 (such as *Igf1* and *Fgfbp1*) of the 82 FG.1 genes were shared by MG.3, but FG.1 had almost no overlapping featured genes with MG.1 or MG.2 (Fig. 5d). Other FG subpopulations' featured genes did not display overlap with MG.3 genes (Supplementary information, Fig. S4b). These findings suggested that FG.1 cells exhibited characteristics of dorsal cells. Interestingly, along the A–P axis, the location of each cell population on the FDL plot reflects its actual anatomical position in the endoderm (Fig. 5a, b).

Along each A–P line, we identified distinct gene clusters with region-specific expression patterns (Supplementary information, Fig. S4c–e, Table S3). Through Reactome and KEGG pathway analyses, we identified a number of signaling pathways enriched along the A–P lines (Fig. 5e–g; Supplementary information, Table S3). RA is essential for establishing the FG–HG boundary^{13,16,45}. In our analysis, we found that RA signaling was mainly enriched in the MG along line-I, indicating the important role of RA in establishing this segment of the endoderm. Fibroblast growth factor receptor (FGFR) signaling components, which regulate lip cell differentiation,^{46–48} were found to be highly expressed in the AL region along line-II. Consistent with previous findings that Wnt signaling is important for maintenance of HG identity,^{11,12,17,46} we observed that Wnt signaling was enriched in HG.1 and HG.2 cells in line-I and line-III, respectively. These findings indicate that the signaling pathways enriched in each line may regulate the formation of endoderm segments, although the roles of other newly identified enriched pathways, such as Notch, Hippo, and Hedgehog (Fig. 5e–g), and their interactions in regulating the A–P segment formation require further investigation.

By comparing cells located in line-II and cells at the same position along the A–P axis in line-I, we identified two gene clusters and several signaling pathways involved in dorsal/lateral vs ventral midline differentiation (Fig. 5h; Supplementary information, Fig. S4f, Table S3). Differentiation of HG.1 and HG.2 also represents D–V differentiation (Fig. 4f; Supplementary information, Fig. S4g). Similarly, by comparing cells located in line-III (MG.3) and cells at the same A–P position in line-I (MG.1 and MG.2), we found that two gene clusters were differentially expressed between the medial and lateral regions (Fig. 5i; Supplementary information, Fig. S4h, Table S3). Most of the signaling pathways were enriched in the medial region (Fig. 5i; Supplementary information, Table S3), suggesting that MG.3 differentiation was strikingly regulated by local signals. Because cells dissected from the left and right regions of the AL and MG were nearly identical at the transcriptomic level (Figs. 3b and 4b), we speculated that at the 9-SS, asymmetry between the left and right sides of the endoderm has not yet been established.

Hedgehog pathway genes were enriched in the dorsal region of the FG along the dorsal/lateral–ventral midline (D/L–V.M.) axis

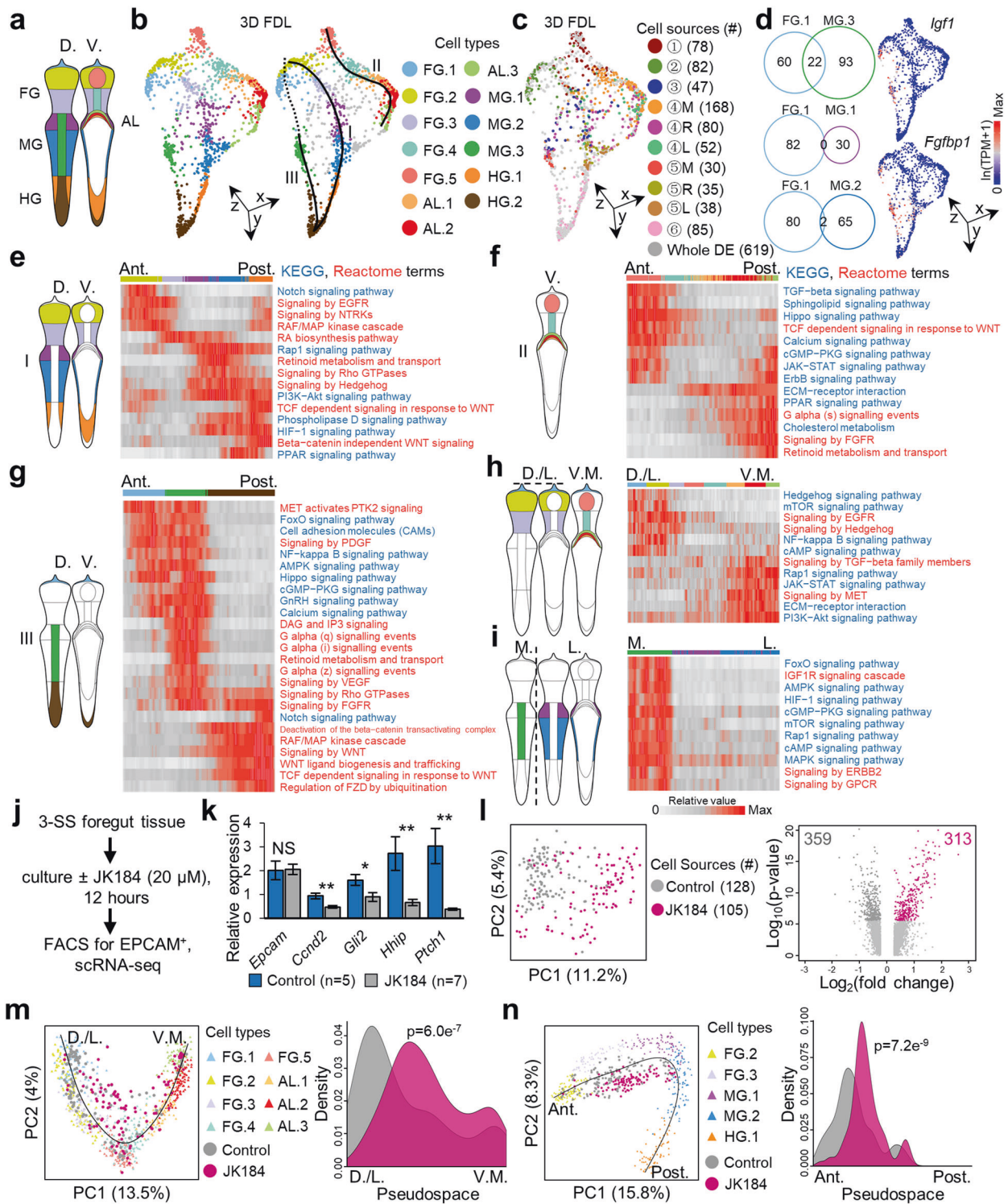
(Fig. 5h). To verify the roles of the Hedgehog pathway in formation of endoderm axes, we treated 3-SS FG endoderm tissues with a Hedgehog inhibitor, JK184⁴⁹ (Fig. 5j–n; Supplementary information, Fig. S4j). Notably, in the FG, the Hedgehog pathway is also highly expressed in the anterior end along the A–P axis (line-I) (Supplementary information, Fig. S4i). After 12 h of culture, the dorsal FG markers *Ccnd2*, *Gli2*, *Hhip*, and *Ptch1*, but not the endoderm marker *Epcam*, were downregulated in JK184-treated FG explants (Fig. 5k). We further performed scRNA-seq of the sorted EPCAM⁺ cells from the explants. Principal component analysis (PCA) and differential gene expression analyses revealed that JK184-treated endoderm cells were significantly different from control cells (Fig. 5l). Moreover, PCA and pseudospace analyses showed that JK184 treatment led to a shift of FG cells to the ventral and posterior axes (Fig. 5m, n). Taken together, these data indicate that the Hedgehog pathway plays a role in regulating the formation of the A–P and D/L–V.M. axes of the endoderm.

In summary, our findings revealed the layout of cell subpopulations along 3D axes of the endoderm germ layer and identified many novel but functionally uncharacterized signaling pathways and gene sets enriched during endoderm patterning.

Temporal development of DE patterning

To understand temporal lineage differentiation during endoderm patterning, we performed scRNA-seq on 1121 endoderm cells at the 3-SS or 6-SS (Fig. 6a; Supplementary information, Table S1). After quality control and removal of nonendoderm lineages (Supplementary information, Fig. S1c–e), 315 endoderm cells at the 3-SS and 535 cells at the 6-SS were retained for further analyses. GCN clustering analysis identified two major cell populations in the 3-SS endoderm (Fig. 6b; Supplementary information, Fig. S5a, Table S2). Based on the expression of marker genes and ISH results, we confirmed that one population, which expressed *Pax9*, was an FG population (Fig. 6c, d; Supplementary information, Fig. S5b). Although a fraction of cells in the AL region began to express *Hhex*, these *Hhex*⁺ cells were not distinct from *Hhex*[−] cells at the transcriptomic level (Fig. 6c; Supplementary information, Fig. S5d). The other population identified in the 3-SS endoderm generally expressed the HG marker *Wnt5b* and the MG marker *Neprn*, suggesting that this population may be the common progenitor of the subsequently differentiated MG and HG (designated mid–hindgut, M–HG) (Fig. 6b–d; Supplementary information, Fig. S5b).

The 6-SS endoderm cells were divided into four populations (Fig. 6b; Supplementary information, Fig. S5a, c, Table S2) based on their transcriptional profiles and the expression patterns of markers of the FG (*Pax9*⁺), AL (*Hhex*⁺), MG (*Neprn*⁺), and HG (*Wnt5b*⁺) (Fig. 6c, d; Supplementary information, Fig. S5c). At this stage, differential gene expression analysis revealed that *Hhex*⁺ cells were significantly different from *Hhex*[−] FG cells (Supplementary information, Fig. S5d). These results indicate that the four major endoderm populations can be recognized in the 6-SS endoderm.



To predict the lineage relationship between the cells at the 3-SS and 6-SS, we conducted velocity analysis⁵⁰ to describe the transcriptional dynamics of splicing kinetics. On the 3D partial least squares discriminant analysis (PLS-DA)⁵¹ plot, we observed that 3-SS FG cells were mainly directed to 6-SS FG and AL cells, and 3-SS M-HG cells were directed to 6-SS MG and HG cells (Fig. 6e). These results are consistent with those of a previous fate-mapping study, which showed that at the early somitogenesis stage, the anterior and posterior cells migrate in a collective

manner and are mainly limited to the anterior and posterior portions of the embryo, respectively.⁵²

To further explore the pathway of specification of the early FG and M-HG, we performed PLS-DA and 3D PCA of the 3-SS FG, 6-SS FG and AL cells. We observed that the 3-SS and 6-SS FG cells clustered together and had similar gene expression profiles, whereas the 6-SS AL cells clustered separately from the FG cells and expressed hundreds of specific genes (Fig. 6f; Supplementary information, Fig. S5c). This analysis identified a developmental

Fig. 5 Mouse endoderm patterning along 3D axes. **a** Schematic representation of the spatial distribution of all mouse endoderm cell subpopulations. **b** The distribution pattern of endoderm cells generated by 3D FDL (left) and the lines of cell distribution along the A-P axis (right). Line-I, FG.2–FG.3–MG.1–MG.2–HG.1; line-II, FG.5–FG.4–AL.1–AL.2–AL.3; line-III, FG.1–MG.3–HG.2. **c** 3D FDL plot showing cell sources. Each dot represents a single cell. Cell counts are indicated in brackets. **d** Venn diagrams showing overlap of FG.1-specific genes with MG.1-, MG.2-, or MG.3-featured genes (left). The cluster-featured genes are shown in Figs. 2c and 4c. Expression levels of the representative FG.1- and MG.3-overlapping genes are shown in the FDL plots (right). **e–i** Schematic representation of the spatial distribution of the cell types (left) and KEGG and Reactome pathway terms (right) enriched in the DEGs along line-I of the A–P axis (**e**), line-II of the A–P axis (**f**), line-III of the A–P axis (**g**), the D./L.–V.M. axis (**h**), and the M–L axis (**i**). Ant., anterior; Post., posterior; D., dorsal; L., lateral; V.M., ventral midline; M., medial. **j** Schematics depicting ex vivo foregut tissue culture. **k** The expression levels of the endoderm marker *Epcam* and the dorsal foregut markers *Ccnd2*, *Gli2*, *Hhip*, and *Ptch1* in cultured foregut endoderm tissues. The data show the means \pm SEM; n = number of independent biological replicates; * P < 0.05; ** P < 0.01; NS, not significant. **l** PCA plot of single-cell transcriptomes from the indicated cells (left). Genes differentially expressed between the control and JK184-treated foregut endoderm cells (right). The numbers indicate significantly highly expressed genes in the control (gray) or JK184-treated (magenta) cells. **m, n** PCA plot (left) and pseudospacial ordering (right) of indicated cells along the D./L.–V.M. axis (**m**) and A–P axis (**n**).

trajectory from 3-SS FG to 6-SS AL cells. We also performed similar analyses on the 3-SS M–HG and 6-SS MG and HG and found that 3-SS M–HG cells were located in an intermediate position between the 6-SS MG and HG cells (Fig. 6g; Supplementary information, Fig. S5b, c), suggesting that 3-SS M–HG cells may follow different paths to differentiate into later MG and HG populations. Moreover, on the FDL plot, 6-SS cells were distributed over the entire 9-SS cell population except for the MG.1 region, but the 3-SS FG and M–HG cells formed two clusters along the A–P axis, overlapping with only a small portion of the 9-SS cells (Supplementary information, Fig. S5e). Interestingly, when 3-SS and 6-SS cells were integrated, we found that FG.1 and MG.3 cells were linked by cells along line-III (Fig. 5b; Supplementary information, Fig. S5f), further indicating the dorsal cell characteristics of 9-SS FG.1 cells. Altogether, these analyses revealed that at the earliest somitogenesis stage, the endoderm layer initially forms two main segments along the A–P axis and are subdivided into four distinctive cell populations and regionalized endoderm.

MG origin of dorsal and ventral PPs

Dorsal and ventral PPs are thought to originate from distinct FG domains.^{22,53} Genetic tracing studies revealed that the majority of ventral PPs originate from the left and right lateral endoderm regions caudal to the AIP,²⁷ where MG.2 is located (Fig. 4d, h). At the same developmental stage, the dorsal PPs are located in the medial region overlapping with the 2nd–4th somites,²⁷ where the cells have specified into MG.3 (Fig. 4d, h). All of these findings indicate that both the dorsal PPs and the majority of ventral PPs are of MG, but not FG, origin. Notably, at the 9–11-SS, the ventral PP domains were identified in the AL area,²⁷ suggesting relocation of MG.1 and MG.2 cells to the AL region. To confirm this relocation, we reanalyzed live whole embryo imaging data at the single-cell level.⁵⁴ At the early somitogenesis stage, we marked the cells located at MG.1–3 regions and tracked their subsequent movement trajectory during development. We observed that as the AL extended caudally, the marked anterior lateral cells moved ventrally into the AL region (Fig. 6h; Supplementary information, Fig. S5g), consistent with findings from studies of chick gut morphogenesis,⁵⁵ and MG.3 cells remained at the medial region (Fig. 6h; Supplementary information, Fig. S5g). Curiously, on the FDL plot, we clearly observed a path connecting the MG.2 and AL.3 groups, and the latter group sporadically expressed the pancreatic marker gene *Pdx1* (Figs. 3c and 6i; Supplementary information, Fig. S5h). This finding indicates that during the relocation from MG.2 to AL.3, the fate of MG.2 cells changes accordingly. Therefore, these analyses, combined with those of previous lineage tracing studies,^{27,43} revealed that both dorsal PPs and the majority of ventral PPs are originally derived from MG cells.

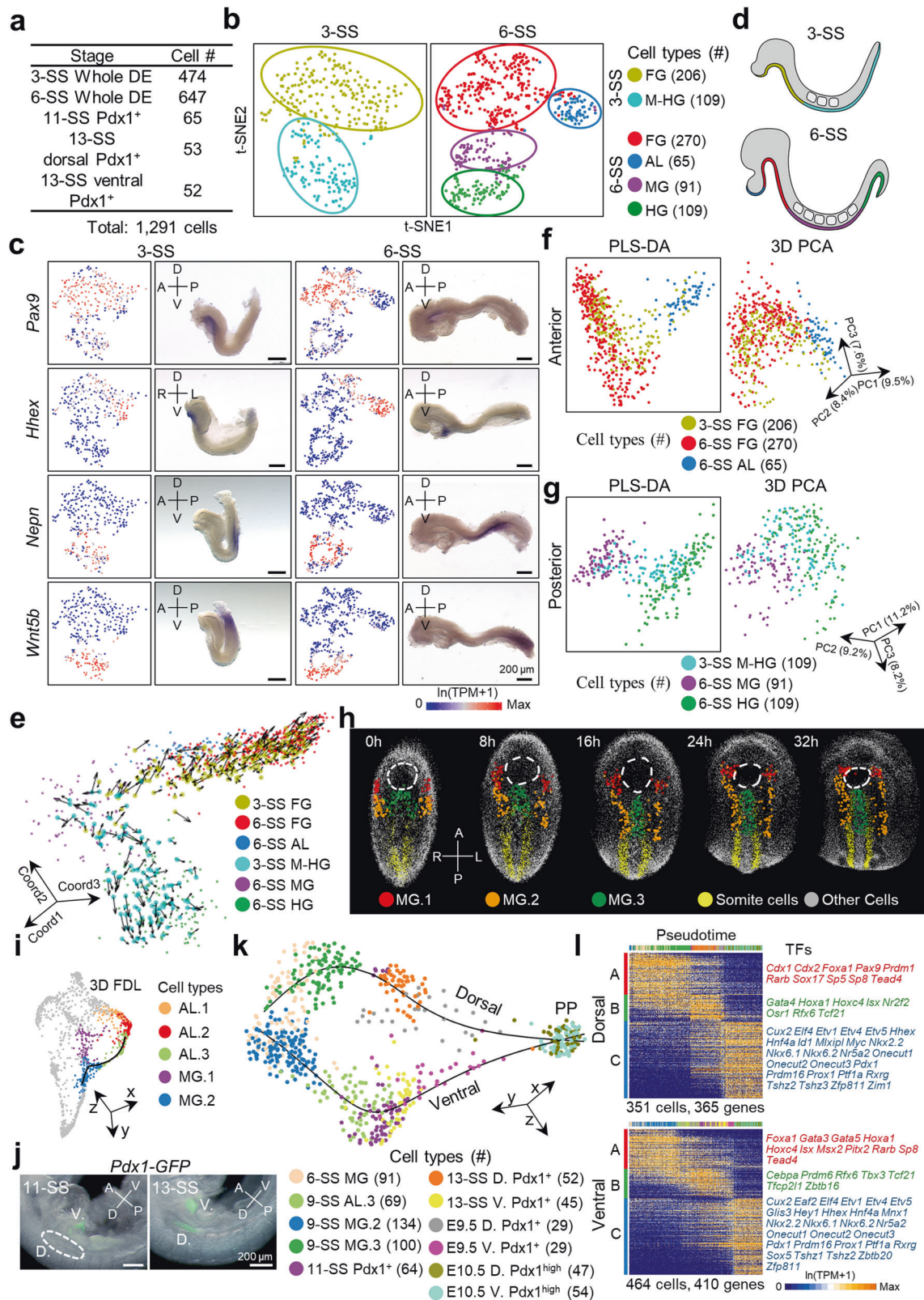
To map the entire developmental path of early PPs, we used a *Pdx1-GFP* transgenic mouse line.⁵⁶ We sorted *Pdx1-GFP*⁺ cells from roughly dissected tissue containing PP cells at the 11-SS, and

separately from the dorsal and ventral *Pdx1*-expressing regions at the 13-SS (at this stage, dorsal vs ventral *Pdx1*⁺ regions can be distinguished under a dissection microscope) (Fig. 6a, j; Supplementary information, Table S1). Our previous study included scRNA-seq of dorsal and ventral E9.5 *Pdx1-GFP*⁺ cells and differentiated PP cells with high *Pdx1-GFP* expression at E10.5.²³ We combined the datasets of the previously described *Pdx1-GFP*⁺ cells with those from 6-SS MG cells and 9-SS MG.2, MG.3, and AL.3 cells and performed 3D FDL to define the dorsal and ventral PP cell differentiation pathways. Both pathways started from 6-SS MG cells and separately follow a dorsal path of 6-SS MG–MG.3–early dorsal *Pdx1*⁺–PP or a ventral path of 6-SS MG–MG.2–AL.3–early ventral *Pdx1*⁺–PP (Fig. 6k). Notably, almost all *Pdx1*⁺ cells at the 11-SS were associated with the ventral path, indicating that *Pdx1*-expressing cells in the ventral region differentiate prior to those in the dorsal region. We also identified different sets of genes and signaling pathways involved in these two pathways (Fig. 6l; Supplementary information, Fig. S5i, Table S4). Altogether, these experiments, together with our previous study,²³ revealed that common MG cells follow distinct ventral vs dorsal pancreatic differentiation pathways to eventually develop into the same cell type.

Cell types in human endoderm

To investigate whether the endoderm patterning and axis paths observed in mice is conserved in humans, we performed scRNA-seq of FACS-sorted EpCAM⁺ cells from the endoderm tissues of 7 human embryos ranging from the 8–15-SS (Fig. 7a; Supplementary information, Fig. S6a, Table S1). In humans, the gut tube morphology at these stages is similar to that at the 9-SS in mice. For one 12-SS embryo, to retain the spatial information, we dissected cells from the AIP (⊙), the segment between the AIP and CIP (⊙), and the CIP (⊙) for scRNA-seq (Fig. 7b; Supplementary information, Table S1). To quickly complete construction of cDNA libraries and avoid sample loss, we used a relatively high-throughput mSTRT-seq method,³⁵ which was similar to the Smart-seq2 protocol used to generate cDNA. However, mSTRT-seq involves unique barcoding and sample pooling during library construction to overcome the low-throughput limitation of Smart-seq2. Although this method generated only 3' transcripts, we were able to detect 4000–7000 genes per cell (Supplementary information, Fig. S6b, c). In total, 2781 cells passed the quality control step; however, after removing other cell types based on marker gene expression, 1544 cells were identified as endoderm cells, indicating that ~50% of the sorted EpCAM⁺ cells in human embryos were DE cells (Supplementary information, Fig. S6d, e).

Similar to our analyses of the mouse endoderm, GCN clustering and marker gene expression analyses identified four major endoderm cell types (the human (h) FG (hFG) (*PAX9*⁺), hAL (*HHEX*⁺), hMG (*ISX*⁺), and hHG (*CDX2*⁺)) at the 8–15-SS (Fig. 7c–h; Supplementary information, Table S5). Consistently, cells dissected



from the AIP (region ①) were predominantly in the hFG and hAL, with a small fraction located in the hMG, while those dissected from regions ② and ③ were found in the hMG and hHG, respectively (Fig. 7f). On the t-SNE plot, cells from each SS intermingled, indicating that during this developmental period,

the endoderm cell fate has not undergone dramatic changes (Fig. 7e).

The second round of GCN clustering analysis identified five hFG cell clusters (hFG.1–5), three hMG clusters (hMG.1–3), and two hHG clusters (hHG.1–2) (Fig. 7i; Supplementary information,

Fig. 6 Differentiation pathways of the endoderm and pancreas. **a** Overview of E8.25 (3-SS/6-SS) endoderm cells and E8.75 (11-SS/13-SS) Pdx1-GFP⁺ single cells sequenced in this study. **b** The t-SNE plot shows distinct endoderm cell populations from 3-SS (left) or 6-SS (right) embryos. Each dot represents a single cell. Cell counts are indicated in brackets. **c** Endoderm population marker genes are shown on the t-SNE plot (left) and were validated by whole-mount ISH of 3-SS and 6-SS embryos (right). $n > 3$ for each gene. Scale bars, 200 μ m. **d** Schematic representation of the spatial distribution of endoderm cell populations at the 3-SS and 6-SS. The cell types are colored as shown in **b**. **e** RNA velocities of 3-SS endoderm cells are visualized on the 3D PLS-DA plot of 6-SS endoderm cells. Arrows point to future cell states. **f, g** The differentiation pathways of anterior (**f**) or posterior (**g**) parts of the endoderm analyzed by PLS-DA and 3D PCA. **h** Migration of marked cells on the endoderm layer. These images were obtained by reanalyzing published data.⁵⁴ The white dotted circles indicate the AIP. Gray dots indicate individual cells. Red, orange, and green dots indicate individual MG.1, MG.2, and MG.3 cells, respectively. Yellow dots indicate the positions of somatic cells. The dark area represents no cell background. 0 h, time point zero of this reanalysis. **i** The line indicates the cells located between the AL and MG populations on the 3D FDL plot. **j** Pdx1-expressing regions dissected from 11-SS (left) and 13-SS (right) embryos. V, ventral; D, dorsal; A, anterior; P, posterior. Scale bars, 200 μ m. **k** 3D FDL of endoderm cells and PPs from embryos at different stages. The E9.5 Pdx1-GFP⁺ (Pdx1⁺), E10.5 dorsal Pdx1-GFP^{high}, and ventral Pdx1-GFP^{high} cells were obtained from published resources.²³ The black curves indicate pathways of dorsal and ventral pancreatic lineage differentiation. **l** Heatmap of DEGs during dorsal (upper) and ventral (lower) pancreatic lineage differentiation. TFs are listed on the right.

Fig. S6f–l, Table S5). However, the hAL cells were relatively homogeneous (Fig. 7i; Supplementary information, Fig. S6i, Table S5). To understand the similarity between human and mouse (m) endoderm cell subpopulations, we performed unsupervised canonical component analysis (CCA)⁵⁷ to align the human and mouse datasets using orthologues annotated by Ensembl genome annotation system (<http://www.ensembl.org/index.html>). The top 10 canonical components were used to calculate the similarity of cell subpopulations between humans and mice; species-specific genes were ignored. Interestingly, each subpopulation in the human endoderm was found to correspond to a subpopulation in the mouse endoderm, except hFG.4, which was more similar to mFG.5 than to mFG.4 (Fig. 7j). Curiously, the homogeneous hAL population at the 8–15-SS was most similar to mAL.3 (Fig. 7j). The similarity of cell types between humans and mice indicates that endoderm patterning is evolutionarily conserved in these two species.

To confirm that no human cell type was omitted due to use of a well-based mSTRT-seq method, we dissected endoderm tissue from a 14-SS human embryo and performed scRNA-seq using the 10 \times Genomics platform. After filtering poor-quality cells and other cell types based on marker gene expressions, 2152 cells were identified as endoderm cells, with ~2000–3000 genes in each cell (Supplementary information, Fig. S7a–f, Table S1). We used the MNN algorithm to correct for methodological batch effects and projected the mSTRT-seq cells onto a 10 \times Genomics plot. We found that the eleven cell populations identified by mSTRT-seq were distributed throughout the t-SNE and 3D FDL plots of the 10 \times Genomics data (Supplementary information, Fig. S7g, h), indicating that no cell type was undetected by the mSTRT-seq. Therefore, we identified the major cell populations and their subpopulations in human endoderm.

Patterning of 3D axes in the human endoderm

We performed 3D FDL analysis to resolve the axial distribution of human endoderm cells. On the FDL plot, hMG.2 formed two small branches that connected hMG.1 and hMG.3 (Fig. 8a), indicating the cellular heterogeneity of hMG.2 population. Hierarchical clustering using hMG.1- and hMG.3-related genes identified two subgroups (hMG.2.1 and hMG.2.2) in hMG.2 cells, which expressed 10 hMG.1-related genes and 8 hMG.3-related genes (Supplementary information, Fig. S8a–c), respectively.

Similar to our previous analysis in mice, we identified three lines of cell ordering along the A–P axis (Fig. 8a), which proceeded through the cell populations hFG.2–hFG.3–hMG.1–hMG.2.1–hHG.1 (line-I), hFG.5–hFG.4–hAL (line-II), and hFG.1–hMG.3–hMG.2.2–hHG.2 (line-III). Generally, the cell distribution patterns of line-I and line-II in the human endoderm were similar to those in mice, whereas the two species displayed obvious differences in line-III. Specifically, along human line-III, hFG.1 and hMG.3 were directly linked, but hMG.3 was not directly connected to hHG.2 and was separated by hMG.2 cells

(Fig. 8a). Based on this information, and referring to the patterns observed in mice, the relative position of cell populations along the A–P axis on the FDL plot, and the gut tube shape revealed by a 3D atlas study of human embryos,⁵⁸ we predicted a model of cell distribution patterns on the human endoderm (Fig. 8b). The ventral portion of this model is similar to that of mice; however, in the dorsal portion, hFG.1 and hMG.3 are located on the dorsal side and extend to the position of the hAL on the A–P axis. The hMG.2 segment is located between the hMG.1/hMG.3 and hHG segments (Fig. 8b).

The distributions of these cell populations were verified by immunofluorescence against the line-I/II-specific gene *ISL1* and the line-III-specific gene *SLIT2* (Fig. 8c). Notably, both of these genes were not expressed in hMG.2 cells. On the cross and sagittal sections of the hFG and hHG positions, we observed that *ISL1*⁺ cells and *SLIT2*⁺ cells were indeed located at the ventral and dorsal side, respectively (Fig. 8c, d; Supplementary information, Fig. S8d). However, in the mouse embryo, *ISL1* and *SLIT2* were expressed in both dorsal and ventral FG regions (Fig. 8d, e). Similar to our observations in mice, the ventral hFG is thickened into a stratified layer (Fig. 8c; Supplementary information, Fig. S8d). To confirm whether differentiation of hFG.1 cells occurs at an earlier developmental stage, we performed scRNA-seq on endoderm cells from one 2-SS, one 4-SS, and one 5-SS human embryo (Supplementary information, Fig. S8e, f, Table S1). We performed 3D FDL analysis of the single cells of the 2–15-SS embryos and found that hFG.1 cells had been generated at all of those stages (Fig. 8f).

Similar to the analyses in mice, along each A–P line, we identified region-specific gene expression patterns and signaling pathways through Reactome and KEGG pathway analyses and identified both conserved or species-specific pathways (Fig. 8g–i; Supplementary information, Fig. S8g–i, Table S6). We also compared the differences between cell populations along the D–V axis, including hFG.1 (dorsal) vs hFG.4/hFG.5 (ventral), and hHG.1 (ventral) vs hHG.2 (dorsal) populations (Fig. 8b). Interestingly, along the D–V axis, many signaling pathways are enriched in dorsal hFG.1 and hHG.2 cells, and some pathways, such as the Hippo, JAK-STAT, and Wnt pathways, are enriched in both hFG.1 and hHG.2 populations (Fig. 8j, k; Supplementary information, Figs. S6l, S8j, Table S6).

Therefore, we revealed the spatial patterning of the human DE and discovered similarities and differences in the arrangement of subpopulations between humans and mice.

DISCUSSION

Understanding DE patterning is critical for understanding the regulatory mechanisms of visceral organogenesis. Traditional approaches, such as genetic studies, gene expression analysis of bulk cells, and immunostaining of limited markers, have been

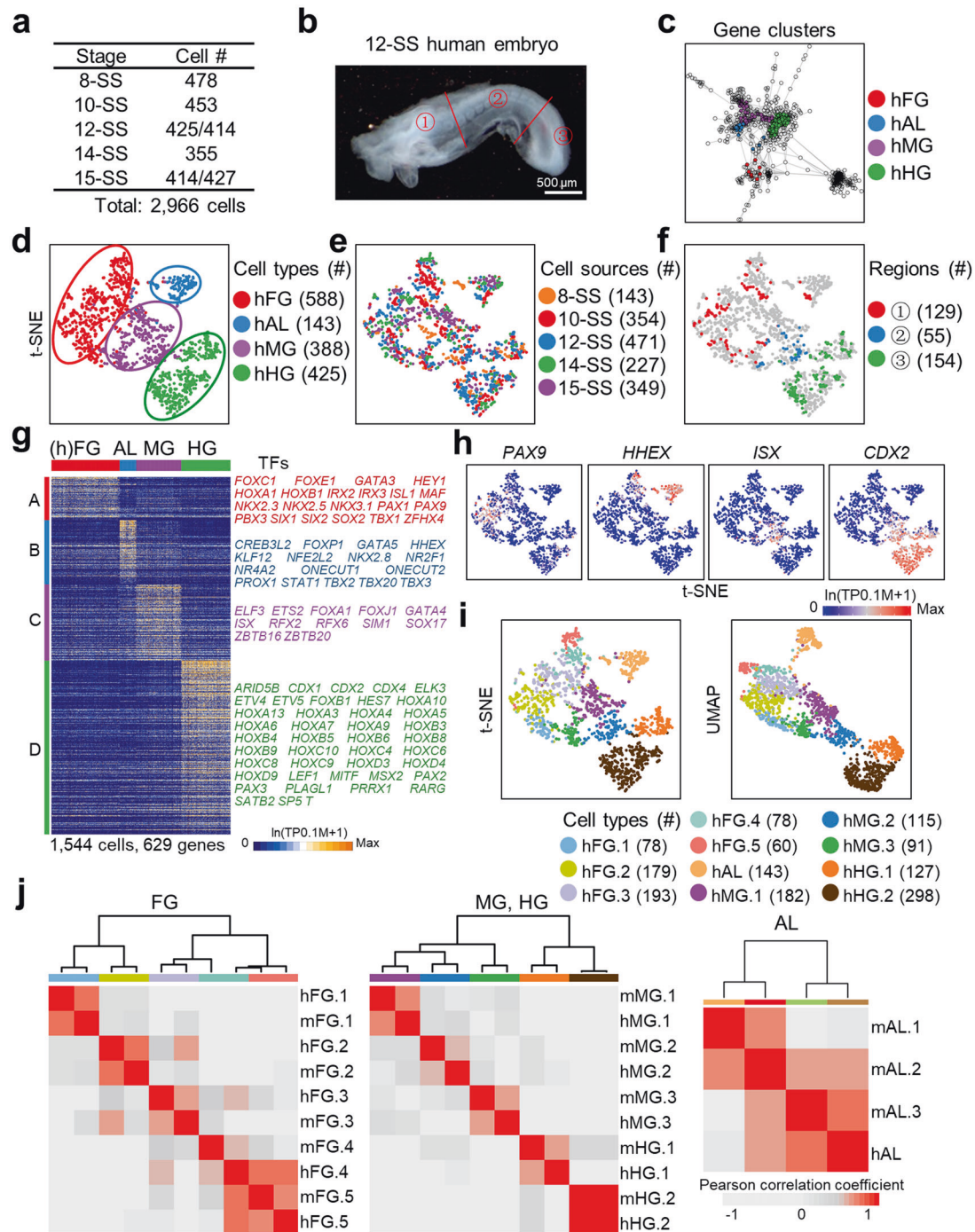


Fig. 7 Identification of human endoderm cell types. **a** Overview of 2966 human embryonic cells analyzed in this study. The numbers show the cell counts from the indicated embryos at different developmental stages. **b** Schematic representation of dissected endoderm regions in a 12-SS human embryo. Scale bar, 500 μ m. **c** Distinct gene clusters differentially expressed in 8–15-SS human endoderm cells. Each dot represents a GCN-associated gene. The cell type-specific gene clusters are colored. **d, e** The t-SNE plot shows four distinct cell types (**d**) and cell sources (**e**) in the 8–15-SS human endoderm. Each dot represents a single cell. Cell counts are indicated in brackets. **f** Distribution of cells from the dissected regions in **b**. **g** Heatmap of 629 DEGs identified four distinct endoderm cell populations and four groups (A–D) of genes. Each column represents a single cell, and each row represents one gene. Group-specific TFs are listed on the right. **h** Expression levels of marker genes are shown on the t-SNE plots. **i** The t-SNE (left) and UMAP (right) plots show 11 distinct cell subpopulations from the 8–15-SS human endoderm cells. The positions of the cell types in **d** and **i** were inferred according to marker gene expression and spatial information of mouse endoderm cells. **j** Similarity analyses of human and mouse FG (left), MG, HG (middle), and AL (right) cells. The color scale represents the Pearson correlation coefficient.

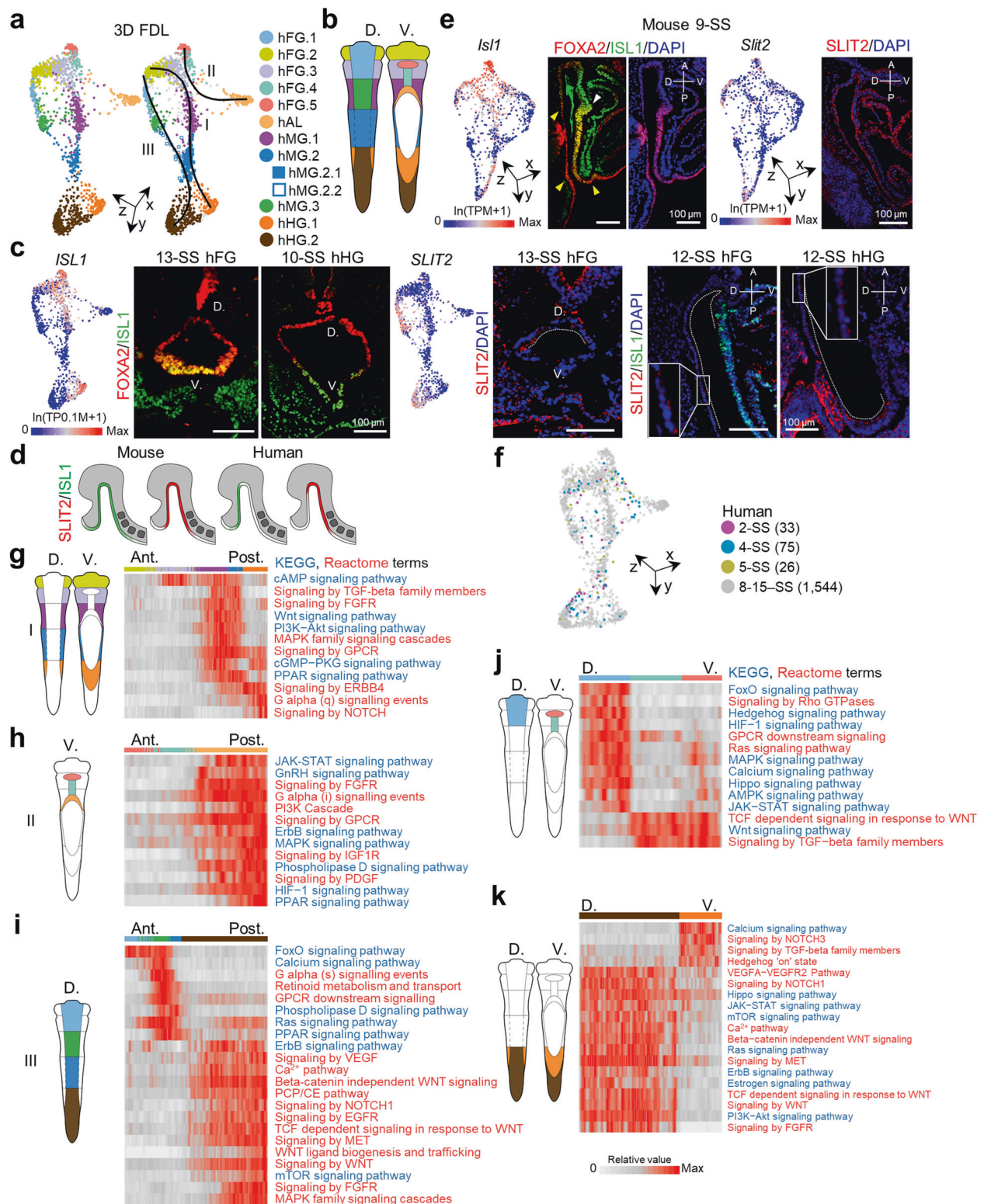


Fig. 8 Human endoderm patterning along 3D axes. **a** Distribution pattern of endoderm cells generated by 3D FDL (left) and lines of cell distribution along the A-P axis (right). Line-I, hFG.2–hFG.3–hMG.1–hMG.2.1–hHG.1; line-II, hFG.5–hFG.4–hAL; line-III, hFG.1–hMG.3–hMG.2.2–hHG.2. A dot within the FDL plot represents a single cell. **b** Schematics showing the spatial distribution of human endoderm cell subpopulations. **c** The expression levels of *ISL1* and *SLIT2* are shown on the FDL plots (left). Validation by immunofluorescence of human embryo sections is shown in the right panels. The *SLIT2* signal is indicated by a dotted line. Scale bars, 100 μ m. **d** Schematic representation of *ISL1* and *SLIT2* expression patterns in mice and humans. **e** FDL plots (left) and immunofluorescence (right) showing the expression of *ISL1* and *SLIT2* in the mouse FG. The white arrowhead indicates a region of higher *ISL1* expression. Yellow arrowheads indicate regions of lower *ISL1* expression. $n > 3$. Scale bars, 100 μ m. **f** 3D FDL plots showing the distribution pattern of various SS endoderm cells. **g–k** Schematic representation of the spatial distribution of the cell types (left) and KEGG and Reactome pathway terms enriched in the DEGs (right) along line-I of the A-P axis (**g**), line-II of the A-P axis (**h**), line-III of the A-P axis (**i**), and the D-V axis (**j**, **k**).

employed to describe the landscape of the endoderm sheet and the regulation of patterning along different axes.^{2,59} However, these methods cannot determine the cell types/subtypes involved in this patterning and their precise layout or the intrinsic connections between cell lineages. Recent studies employing scRNA-seq on the 10× Genomics platform attempted to resolve the endoderm map in mice, and several cell clusters along the A–P axis were identified;^{31,32} however, due to high noise and low sensitivity for low-abundance transcripts (Supplementary information, Fig. S2b, c, g–i), this method has limited application in defining cell types with only subtle differences at the transcriptomic level, and thus the subpopulations, precise patterning of the DE and molecular features and signaling pathways along the 3D body axes have not been resolved. In this study, we used a more sensitive method, Smart-seq2, combined with GCN clustering analysis and other algorithms to delineate endoderm cell subtypes and their distributions on the A–P, D–V, and M–L axes. These findings were carefully verified by ISH and referenced spatial information provided by microdissection.

At the earliest somite stage, the mouse endoderm formed two segments, FG and M–HG, which quickly differentiated into four major cell types by 6–SS. Our analysis clearly demonstrated that AL is a major cell population independent of the FG, MG, and HG populations. These major populations were further divided into 13 subpopulations, which constituted the DE landscape. Throughout our study of this landscape, we observed a phenomenon of “midline specification”; that is, the mFG.5–mFG.4–mAL cells formed a ventral midline, whereas mMG.3 is located on the dorsal midline. Cells from mFG.5 express the thyroid marker *Nkx2.1*,⁶⁰ and lineage tracing revealed that mAL cells generate the liver and ventral pancreas and that mMG.3 cells are the progenitors of the dorsal pancreas.^{27,43} We therefore hypothesize this “midline specification” is the molecular basis for the generation of the organs associated with the gut tube.

Midline specification also resulted in three cell distribution lines along the A–P axis. Each line involves different cell signaling pathways, many of which are functionally uncharacterized in endoderm differentiation. A clear distinction between these three lines is critical to our future studies of cell signaling mechanisms in regulation of A–P patterning. Curiously, although mFG.1 and the other line-III cell populations were not directly linked (Fig. 5b), mFG.1 expressed many line-III-specific genes and was therefore considered to be a line-III cell population (Fig. 5b, d; Supplementary information, Fig. S5f). Using our precise cell distribution map, we could also accurately compare the differences in gene expression between cells on the D–V and M–L axes.

Understanding DE patterning also enhanced our understanding of the origin of endoderm-derived organs. PPs are generally considered to have both the dorsal FG and ventral FG origins. However, based on the information of DE regions identified by scRNA-seq, combined with lineage tracing studies^{27,43} and live imaging data,⁵⁴ we conclude that dorsal PPs and the majority of ventral PPs originate from MG cells. With the scRNA-seq data of the relevant DE cells and the late differentiated *Pdx1*⁺ cells, we deciphered a comprehensive PP differentiation pathway, where both dorsal and ventral PPs originate from MG cells and follow separate paths to generate the same cell type. This analysis also clarified that *Pdx1*⁺ cell differentiation in the ventral domain occurs prior to that in the dorsal domain, however, the newly generated ventral *Pdx1*⁺ cells also retain the potential to develop into hepatoblasts and extrahepatic bile ducts.^{23,61} In addition to the MG domain, the midline region of the AL also contributes a small fraction of PPs.²⁷ This domain was presumably derived from the FG, given that mFG.4 cells are located on the ventral midline adjacent to mAL.1 and the cells at the ventral midline caudally migrate to the AL in chicks and mice.^{52,62}

We performed similar endoderm patterning analyses on human embryos. At the 8–15–SS, human endoderm could also be classified into four major populations and further divided into subpopulations. Interestingly, each human subpopulation has its own corresponding subpopulation(s) in mice (Fig. 7j), indicating that endoderm development is generally evolutionarily conserved between these two species. However, in humans, we observed “midline specification” in the dorsal FG but not the hMG.2 segment. This difference needs to be studied in the context of subsequent developmental processes to understand its functional significance.

Endoderm patterning is regulated by signals secreted from other germ layers, such as the neuroectoderm and cardiac mesoderm.^{2,59} Such signals are expected to play important regulatory roles in the midline specification process. Understanding how different cellular signals (extrinsic cues) coordinate with intrinsic factors to regulate precise cell fate differentiation and pattern formation should be the focus of future mechanistic studies. Induction of proper endoderm progenitor cells from hPSCs in vitro is also a critical step for endoderm-derived organ regeneration. For example, induction of pancreatic endocrine cells in vitro first requires induction of stem cells into MG.3 endoderm progenitor cells, which are the proper progenitors of dorsal pancreatic lineages. Therefore, our work provides a blueprint for researchers to evaluate the quality of induced organ progenitor-specific endoderm cells.

MATERIALS AND METHODS

Mouse lines

B6C3F1 (the F1 hybrid of female C57BL/6 and male C3H) females were crossed with the B6C3F1 and *Pdx1-GFP*⁵⁶ males to generate embryos used in this study. Appearance of the vaginal plug was counted as E0.5. All mice were bred and maintained under pathogen-free conditions at Peking University. All experiments were conducted following the Animal Protection Guidelines of Peking University.

Human embryos

Human embryos were obtained from elective terminated pregnant women with written informed consent at Haidian Maternal & Child Health Hospital in Beijing, China. The operations were performed in accordance with protocols approved by the Peking University Institutional Review Board (PU-IRB) (Certificate number: IRB00001052-18083).

Endoderm dissection

Endoderm tissue was obtained by trypsinization and microdissection according to a previously published protocol.³⁶ Briefly, after removal of extraembryonic tissue, mouse and human embryos were digested in 1% trypsin in Hank's Balanced Salt Solution (HBSS) for 30 min on ice. Then, 0.25 volumes of fetal bovine serum (FBS) and 0.01% DNase I were added to stop the reaction and digest genomic DNA. The endoderm and ectoderm were carefully separated using fine tip forceps or a microsurgical knife. For the AL, the ventral portion at the position of the first two somites was dissected, and then the tissue was divided into three equal parts (Supplementary information, Fig. S1a). For the MG, we first dissected the endoderm tissue at the position of the 3rd–9th somite segment, and then dissected the medial and lateral tissues according to the lines shown in Supplementary information, Fig. S1b. For each batch, we dissected 30–80 embryos from 3–5 mice, and only 3–SS, 6–SS, or 9–SS embryos were used for endoderm dissection. On average, ~10–15 embryos at the same somite stage in each batch were used for single-cell experiments.

Dorsal and ventral pancreatic endoderm tissues from E8.75 *Pdx1-GFP* embryos were dissected under a fluorescence stereomicroscope (Zeiss, Lumar.V12).

Cell sorting

Dissected tissue was dissociated by incubation in 0.25% trypsin for 5 min at 37 °C. Then, 0.4 volumes of FBS was added to stop the reaction. APC anti-human CD326 (EpCAM) (Biolegend, Cat# 324208) and APC anti-mouse CD326 (EpCAM) (Thermo Fisher Scientific, 17-5791-82) antibodies were used for cell sorting. Cells were sorted into 96-well plates using a FACS Aria SORP flow cytometer (BD Biosciences).

scRNA-seq library preparation

RNA-seq libraries were prepared according to the Smart-seq2,³⁰ mSTRT-seq,³⁵ or droplet-based methods. For Smart-seq2, 2 ng of cDNA was used to prepare libraries using a TruePrep DNA Library Prep Kit (Vazyme Biotech, TD502). For mSTRT-seq, cDNAs from 48–96 different single cells were pooled together. Four-cycle PCR using biotinylated index primers introduced a biotin tag to the 3' end of the cDNA. After cDNA fragmentation using a Bioruptor Plus sonicator (Diagenode), the 3' terminal cDNA segments were captured using Dynabeads MyOne Streptavidin C1 (Thermo Fisher Scientific, 65002). Thirty nanograms of cDNA was used to construct libraries using a Kapa Hyper Prep Kit (Kapa Biosystems, KK8505). For the droplet-based method, cDNA synthesis and library construction were performed using the Single Cell 3' Reagent Kit v2 (10x Genomics) following the user guide.

ISH

The cDNA templates for riboprobes were generated using SuperScript II Reverse Transcriptase (Thermo Fisher Scientific, Cat# 18064071) from total RNAs isolated from E8.5 embryos. Primers for probe templates (Supplementary information, Table S7) contained the T7 or SP6 RNA polymerase promoter attached to the 5' end of the gene-specific sequences. Amplified riboprobe templates were cloned into the pTOPO-Blunt vector (Mei5 Biotechnology, MF021) and validated by sequencing. The riboprobes were generated using a DIG RNA Labeling Kit (Roche, Cat# 11277073910).

Whole-mount ISH was performed as previously reported.⁶³ Briefly, embryos were fixed in 4% paraformaldehyde (PFA) at 4 °C overnight, washed with PBS with 0.1% Tween-20 (PBST), and dehydrated through a methanol series. Antisense riboprobes were used for hybridization, and BM Purple was used for detection. Images were acquired using a Zeiss Lumar.V12 or Leica M205 FCA microscope. After whole-mount ISH, embryos were dehydrated through a methanol series and embedded in paraffin. The 7-μm-thick sections were dewaxed, rehydrated, and counterstained with eosin. Images were acquired with a Zeiss Imager.M2 microscope.

Whole-mount fluorescent ISH (FISH) was performed as previously described.^{64,65} Briefly, embryos were fixed in 4% PFA at 4 °C overnight, washed with PBST, and dehydrated through a methanol series. *Upk3a* (RNAscope, 505891) and *Ghrl* (RNAscope, 415301-C2) probes were used for hybridization following the manufacturer's instructions of the RNAscope Multiplex Fluorescent Reagent Kit v.2. Embryos were imaged in PBST using the Leica TCS SP8.

Immunofluorescence

Human embryos were fixed overnight at room temperature in 10% neutral buffered formalin (NBF). After fixation, embryos were washed 3 times in PBS, dehydrated, embedded in paraffin, and sectioned into 5-μm-thick sections. Sections were deparaffinized and dehydrated. After washing and blocking (20% FBS in TBS with 0.05% Tween-20), anti-ISL1 IgG (DSHB, 39.4D5), anti-FOXA2 IgG (SEVEN HILLS, WRAB-FOXA2), anti-SLIT2 IgG (Novus Biologicals, NBP1-80742, NBP2-20398), Alexa Fluor 594 donkey anti-rabbit IgG (Thermo Fisher Scientific, A21207), and Alexa Fluor 488 donkey anti-mouse IgG (Thermo Fisher Scientific, A21202) were used for immunofluorescence. Images were acquired using a Zeiss Imager.M2 microscope.

Ex vivo culture

The 3-SS FG endoderm tissues were dissected from E8.25 mouse embryos. After dissection, tissues were transferred to 96-well plates (Corning, Cat# 3599) containing 100 μL culture medium (48.3% Dulbecco's Modified Eagle Medium/Nutrient Mixture F-12 (DMEM/F-12; Sigma, Cat# 11330032), 50% rat serum,⁶⁶ 0.63% glucose, and 1% penicillin/streptomycin) each well and cultured in an incubator (37 °C; 5% CO₂, and 95% air) for 12 h with or without 20 μM of the Hedgehog inhibitor JK184 (Selleck Chemicals, S6565).

For RT-qPCR, the tissues and the detached cells after culturing were collected, and total RNA was prepared using an RNAprep Pure Micro Kit (Tiangen, DP420). cDNA was synthesized using HiScript II Q RT SuperMix for qPCR (+gDNA wiper) (Vazyme Biotech, R223). Primer sequences are listed in Supplementary information, Table S7. For single-cell analyses, the cultured tissues and the detached cells were dissociated into a single cell suspension. EpCAM⁺ cells were sorted for scRNA-seq by mSTRT-seq.

scRNA-seq data preprocessing and alignment

Smart-seq2 data processing. Smart-seq2 scRNA-seq libraries were sequenced as 51-bp single-end reads on an Illumina HiSeq 2500 platform. Read quality was evaluated using FastQC v0.11.3 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Reads were then aligned to the *Mus musculus* reference genome GRCh38/mm10 with TopHat v2.1.0⁶⁷ with the parameters “-o out_dir -G gtf --transcriptome-index bowtie2_index input_fastq”. The aligned reads were counted using HTSeq v0.6.0⁶⁸ with the parameters “-f bam -r pos -s no -a 30”.

mSTRT-seq data processing. mSTRT-seq libraries were sequenced as 150-bp paired-end reads on an Illumina HiSeq 4000 platform. Raw files (fastq format) were split based on cell-specific barcode sequences with a 1-bp mismatch tolerance. PolyA sequences were trimmed from R1 reads (3' end of cDNA). Preprocessed R1 reads from each cell were aligned to the *Homo sapiens* reference genome (GRCh38.p7/hg38) with TopHat V2.1.0 using the previously described parameters. Read quality was evaluated using FastQC v0.11.3 as described for Smart-seq2. Reads were then annotated to genes with featureCounts (v1.5.3)⁶⁹ using the parameters “-a gtf -o out_file -R BAM out_bam”. Thus, gene information was recorded in the bam file with an XT tag. We used SAMtools (v1.3.1) to sort and index the output bam files.⁷⁰ Unique molecular identifiers (UMIs) of each gene were counted with umi_tools (v0.5.0)⁷¹ using the parameters “count --per-gene --gene-tag=XT --method unique -l indexed_sorted_bam -S out_file”.

10x Genomics data processing. The 10x Genomics libraries were sequenced as 150-bp paired-end reads on an Illumina HiSeq 4000 platform. Raw files were processed with Cell Ranger 3.0.1 using the default parameters. Human and mouse reads were respectively mapped to the GRCh38 or mm10 reference genome (version 1.2.0) provided by 10x Genomics. Cells were filtered using Cell Ranger 3.0.1 with default parameters.

Analysis of preprocessed scRNA-seq data

Data preprocessing and filtering. For the Smart-seq2 dataset, read counts were converted to transcripts per million (TPM).⁷² Cells with < 4000 detected gene counts or < 0.3 million reads were regarded as having inferior quality and were removed from further analyses. The TPM values of the remaining cells were imported into Seurat v2.2.1³⁷ and converted to ln-normalized values (lnTPM). To exclude non-endoderm cells, we performed t-SNE followed by spectral density-based clustering (DBSCAN) to identify outliers. Non-endoderm outliers were further clustered and distinguished using the standard Seurat pipeline. Cells with inferior quality and non-endoderm cells were excluded from

further analyses. Additionally, we observed that some endoderm cells that expressed *Trap1a* and other visceral endoderm-related genes experienced a transition from the visceral endoderm to the definite endoderm.³¹ To avoid disturbing downstream dimensional reduction by visceral endoderm-related genes, we identified and tagged the visceral endoderm-derived cells using the following method. A total of 13 genes (*Rhox6*, *Xlr4c*, *Mageb16*, *Cbx7*, *Xlr3a*, *Xlr3c*, *Xlr5a*, *Xlr4b*, *Xlr5b*, *Xlr3b*, *Xlr4a*, *Al662270*, and *Rhox5*), whose Pearson correlations with *Trap1a* were higher than 0.5, were identified as *Trap1a*-related genes. Cells with an average InTPM of these *Trap1a*-related genes > 1.5 were regarded as visceral endoderm-derived cells.

For the mSTRT-seq dataset, gene expression levels were imported into Seurat v2.2.1 and transformed into transcripts per 0.1 million (TP0.1 M), which normalized the number of total transcripts in one single-cell library to 100,000. Then, TP0.1 M was further ln-normalized (lnTP0.1 M). Cells with < 1500 detected genes were removed from further analyses. Non-endoderm cells were identified using t-SNE and cluster determination using the Seurat pipeline. Cells with inferior quality and non-endoderm cells were excluded from further analyses.

Batch effect correction and dimension reduction. Differential expression analysis was performed with the function “FindVariableGenes” in Seurat. Genes involved in the cell cycle and visceral endoderm were excluded by hierarchical clustering. For all mouse single-cell data except that from whole endoderm cells, we used the “mnnCorrect” function in scran v1.10.1⁴⁰ to correct for batch effects. Visceral endoderm-derived and DE-derived cells were also corrected using scran v1.10.1. PCA was performed based on the corrected lnTPM matrix to obtain low-dimension subspace coordinates using the function “RunPCA” in Seurat. t-SNE was performed with the function “RunTSNE” in Seurat based on the pre-executed PCA subspace. Note that batch effect correction was applied only in the dimensional reduction step rather than in the subsequent analyses because the following clustering algorithm can filter out batch effects and other detectable noise.

GCN clustering. First, we defined the co-expression partners of each gene in the range of variably expressed genes of a cell community. We calculated pairwise Pearson correlation coefficients for the variably expressed genes in a pending cell community. Then, the GCN was constructed using the Pearson correlation matrix above a specific threshold (for mouse and human whole endoderm, mouse FG, AL, and HG, the cutoff value was 0.35; for mouse MG and human subpopulations, the cutoff value was 0.3). The genes adjacent to a specific gene in the network were defined as the co-expressed partners (CPs) of this gene. The CP count is equal to the degree of a network. Second, we selected “hub” genes with greater CP counts than the background to build a GCN (for Smart-seq2 dataset, the threshold was 10 CP counts; for the mSTRT-seq dataset, the threshold was 8 CP counts). After matrix sparsification and hub genes selection, the remaining genes and correlation linkages of genes were regarded as vertexes and edges, respectively. The GCN was displayed using FDL in igraph v1.2.2 (<https://cran.r-project.org/web/packages/igraph/>). In the GCN, genes were clustered using hierarchical clustering as co-expressed gene modules (CGMs). In general, each putative cell cluster occupies a unique CGM. Next, batch effects, cell cycle- or visceral endoderm-related CGMs, and genes that were expressed (lnTPM > 1) in >80% of the total cells were regarded as housekeeping genes and excluded. The remaining genes were used for hierarchical clustering to identify the cell groups. Differential expression analysis (see step 4 for details) was performed to identify cell group-specific GCN genes. Differentially expressed GCN-associated genes of the cell groups were colored on the GCN plot. To reduce noise when constructing the GCN, GCN clustering was iteratively performed

twice on the dataset of mouse 9-SS cells and the dataset of human 8–15-SS cells.

Identification of DEGs among the cell populations. To identify DEGs among the cell populations, we used the “FindAllMarkers” function in Seurat. For four major populations of the mouse and human whole endoderm, we retained genes with P value < 10^{-20} and those were expressed in > 10% of the cells in any population but < 80% of the total cells. For subpopulations of a certain major population, the thresholds of DEGs were set to a P value < 10^{-10} (for mouse) or a P value < 10^{-6} (for human), and they were expressed in > 10% of the cells in any subpopulation and < 80% of the major population cells. Next, we performed hierarchical clustering using all DEGs and assigned these genes into several groups according to the expression patterns. Using this approach, we identified DEGs of mono-populations and multi-populations.

Defining body axes. We performed dimension reduction using 3D FDL. An adjacent matrix was constructed with the “BuildSNN” function in Seurat based on the PCA subspace. 3D FDL was performed using igraph. Quantification of the body axes used a principal curve (smoothing spline fitness). Unlike the general principal curve analysis pipeline, we used a predefined curve instead of a random curve to initiate fitting of the principal curve. First, we defined the gravity centers of each putative population involved in a trajectory on the 3D FDL. The gravity centers were connected to generate a line, both ends of which were extended outward (stretch factor = 2) to form a predefined curve. Next, the predefined curve was iteratively fitted using the principal curve algorithm in princurve V2.1.3 (<https://cran.r-project.org/web/packages/princurve/>) using the parameter “smoother = “smooth.spline”, stretch = 2”. The arc length from the beginning of the curve was regarded as the scale of the body axis.

Maximal information-based nonparametric exploration (MINE) of feature-related genes. To evaluate continuous feature-related genes, we used the maximal information coefficient (MIC). The MIC values between the axis scale and each gene were calculated. We selected genes with MIC values > 0.25 and that were expressed in < 80% of the total cells as axis-related genes. MINE was performed using minerva v1.4.7 (<https://cran.r-project.org/web/packages/minerva/>).

Signaling pathway analysis. We combined the KEGG and the Reactome Pathway Database.^{73,74} The average expression of body axis-related and pathway-annotated genes was used as a measure of the strength of the pathways on the body axis.

Analysis of ex vivo cultured cells. First, we used the “mnnCorrect” function in scran v1.10.1⁴⁰ to correct for batch effects between ex vivo cultured and in vivo FG cells. Next, we projected JK184-treated and control cells into the PCA subspaces calculated using FG or line-I-related cells. The pseudospaces of the dorsal/lateral vs ventral midline axis and the A–P axis were quantified using principal curves. In addition, genes that were differentially expressed between JK184-treated and control cells were identified using the “FindAllMarkers” function in Seurat.

PLS-DA. We performed PLS-DA using the region information from the mouse 6-SS dataset. PLS-DA converted gene expression into a linear space, which can better explain the observable variables. Next, early-stage cells were linearly projected into the PLS-DA space to investigate the region development trend. PLS-DA was performed with caret V6.0-80 (<https://cran.r-project.org/web/packages/caret/>).

Defining the differentiation pathway of PPs. We aggregated the cells involved in PP differentiation together. Cell cycle- and visceral

endoderm-related genes were excluded by hierarchical clustering. PC1–20 were used to build an adjacent matrix. Then, 3D FDL was performed as described to define the body axes.

Refining the developmental pseudotimes of the dorsal and ventral cells using dynamic time warping (DTW). Developmental pseudotimes of the related dorsal and ventral cells were separately calculated with smoothing spline fitness based on the 3D FDL subspace. The dorsal and ventral pseudotime-related genes were identified using MINE as described above. Then, the common variable genes for dorsal and ventral *Pdx1*⁺ lineage development were determined. We used DTW based on common genes to align cells on the dorsal and ventral *Pdx1*⁺ cells. Subsequently, pseudotimes of the dorsal and ventral *Pdx1*⁺ cells were refined based on the cell alignment. DTW was performed with dtw V1.20-1 (<https://cran.r-project.org/web/packages/dtw/>). The cross-distance matrix was constructed based on cosine distance. The step pattern matrix was created using “rabinerJuangStepPattern(6, “c”), as follows:

$$g_{ij} = \min\{g_{i-3,j-2} + d_{i-2,j-1} + d_{i-1,j} + d_{ij}, g_{i-1,j-1} + d_{ij}, g_{i-2,j-3} + d_{i-1,j-2} + d_{ij-1}\},$$

where *d* represents elements in the cross-distance matrix and *g* represents elements in the cost matrix.

Defining the surface cells from the endoderm-side in TARDIS-registered embryo A⁵⁴. Published imaging data include the following information: 3D coordinates of cells at different time points, the inheritance relationship of cells in adjacent time points, and the identities of somitic mesoderm and anterior paraxial mesoderm lineages. The images of the cultured mouse embryos were captured from a ventral view; therefore, the endoderm sheet was located on the surface of the embryos. Using smooth surface fitting with Akima v0.6-2 (<https://cran.r-project.org/web/packages/akima/>),⁷⁵ we located this endoderm layer on the digital embryo image. With reference to the cell distribution information obtained through our single-cell analyses and the positions of somitic pairs and the AIP portal on the digital embryo, we digitally marked the MG.1 (red), MG.2 (orange), and MG.3 (green) cells at an earlier SS and traced their migration trajectories during the subsequent culture process. The graphs were generated using rgl v0.100.30.

Alignment of human and mouse endoderm regions. We combined the genes in the human and mouse GCNs to perform unsupervised CCA with the function “RunCCA” in Seurat. The top 10 canonical components were further aligned and refined with the “AlignSubspace” function. The top 10 canonical components were used to calculate the Pearson similarities across human and mouse clusters. We then performed hierarchical clustering to identify closer clusters in the human and mouse endoderms.

Analysis of 10× Genomics preprocessed data. The 10× Genomics dataset was analyzed with the Seurat standard pipeline. We excluded additional poor-quality cells that were missed by Cell Ranger. 10× Genomic data and Smart-seq2/mSTRT-seq data were combined with MNN batch effect correction using scran.

Imputation with pseudocells. We used a pseudocell algorithm to overcome dropouts in our 10× Genomics dataset.⁴¹ First, for each of the four major populations, we randomly sampled 500 pseudocells (for a total of 2000 pseudocells). For each pseudocell, the average expression value of all genes of the 20 nearest neighbor cells in the PC1–PC20 subspace was considered as the gene expression value of the given pseudocell. Differential expression analysis was performed based on four populations of the pseudocells. Next, we performed unsupervised clustering of the pseudocells with the Seurat standard pipeline.

RNA velocity analysis. We performed RNA velocity analysis using the Velocity pipeline with the parameters “run_smartseq2 -o out_dir -e output_name bam_file gtf_file”. Output.loom files were imported into R with the R package velocity. The arrows on the 3D PLS-DA plot were calculated using the function “show.velocity.on.embedding.cor” in velocity.R with the parameters “n=50, scale=‘sqrt’, min.grid.cell.mass=1, grid.n=100”.

DATA AVAILABILITY

The RNA-seq data from this publication have been deposited in the Genome Sequence Archive (GSA) and assigned the identifiers CRA003104 and HRA000280.

ACKNOWLEDGEMENTS

We thank Drs. Ken Zaret, Chris Wright, and Erfei Bi for their critical advice; Drs. Genji Qin and Peng Du for assistance with imaging; Drs. Jing Zhang and Yi Wang for assistance with human embryo collection; members of the Xu laboratory for advice and comments; Ms. Yanzhu Yue and Dr. Aibin He for assistance with the rat serum preparation and tissue culture; and the Peking-Tsinghua Center for Life Science Computing Platform. We thank the flow cytometry Core at National Center for Protein Sciences at Peking University, particularly Ms. Fei Wang and Ms. Yinghua Guo for technical help, and Mr. Ming Du from the Core Facility of the School of Life Sciences for drawing schematics. This work was supported by the National Key R&D Program of China (2019YFA0801500), the National Basic Research Program of China (2015CB942800), the National Natural Science Foundation of China (31521004, 31471358, and 31522036), and funding from Peking-Tsinghua Center for Life Sciences.

AUTHOR CONTRIBUTIONS

C.-R.X. conceived the project; C.-R.X., L.-C.L., and X.W. designed the study; L.-C.L., Z.-R.X., Y.-C.W., Y.F., Liu Y., and Li Y. performed the experiments; L.-C.L., X.W., W.-L.Q., Z.-R.X., X.-X.Y., J.G. and C.-R.X. analyzed the data; L.-C.L., X.W., X.-X.Y., and C.-R.X. wrote the manuscript.

ADDITIONAL INFORMATION

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41422-020-00426-0>.

Competing interests: The authors declare no competing interests.

REFERENCES

- Takebe, T. & Wells, J. M. Organoids by design. *Science* **364**, 956–959 (2019).
- Zorn, A. M. & Wells, J. M. Vertebrate endoderm development and organ formation. *Annu. Rev. Cell Dev. Biol.* **25**, 221–251 (2009).
- Grapin-Botton, A. & Melton, D. A. Endoderm development: from patterning to organogenesis. *Trends Genet.* **16**, 124–130 (2000).
- Lewis, S. L. & Tam, P. P. Definitive endoderm of the mouse embryo: formation, cell fates, and morphogenetic function. *Dev. Dyn.* **235**, 2315–2329 (2006).
- Peng, G. et al. Spatial transcriptome for the molecular annotation of lineage fates and cell identity in mid-gastrula mouse embryo. *Dev. Cell* **36**, 681–697 (2016).
- Peng, G. et al. Molecular architecture of lineage allocation and tissue organization in early mouse embryo. *Nature* **572**, 528–532 (2019).
- Kraus, M. R. & Grapin-Botton, A. Patterning and shaping the endoderm in vivo and in culture. *Curr. Opin. Genet. Dev.* **22**, 347–353 (2012).
- Guio, J. & Jensen, K. B. From definitive endoderm to gut — a process of growth and maturation. *Stem Cells Dev* **24**, 1972–1983 (2015).
- Tam, P. P., Khoo, P. L., Wong, N., Tsang, T. E. & Behringer, R. R. Regionalization of cell fates and cell movement in the endoderm of the mouse gastrula and the impact of loss of *Lhx1*(*Lim1*) function. *Dev. Biol.* **274**, 171–187 (2004).
- Tam, P. P. & Loebe, D. A. Gene function in mouse embryogenesis: get set for gastrulation. *Nat. Rev. Genet.* **8**, 368–381 (2007).
- Dessimoz, J., Opoka, R., Kordich, J. J., Grapin-Botton, A. & Wells, J. M. FGF signaling is necessary for establishing gut tube domains along the anterior-posterior axis in vivo. *Mech. Dev.* **123**, 42–55 (2006).
- McLinn, V. A., Rankin, S. A. & Zorn, A. M. Repression of Wnt/β-catenin signaling in the anterior endoderm is essential for liver and pancreas development. *Development* **134**, 2207–2217 (2007).
- Kumar, M., Jordan, N., Melton, D. & Grapin-Botton, A. Signals from lateral plate mesoderm instruct endoderm toward a pancreatic fate. *Dev. Biol.* **259**, 109–122 (2003).

14. Roberts, D. J. et al. Sonic hedgehog is an endodermal signal inducing Bmp-4 and Hox genes during induction and regionalization of the chick hindgut. *Development* **121**, 3163–3174 (1995).
15. Tiso, N., Filippi, A., Pauls, S., Bortolussi, M. & Argenton, F. BMP signalling regulates anteroposterior endoderm patterning in zebrafish. *Mech. Dev.* **118**, 29–37 (2002).
16. Chen, Y. et al. Retinoic acid signaling is essential for pancreas development and promotes endocrine at the expense of exocrine cell differentiation in *Xenopus*. *Dev. Biol.* **271**, 144–160 (2004).
17. Wells, J. M. & Melton, D. A. Early mouse endoderm is patterned by soluble factors from adjacent germ layers. *Development* **127**, 1563–1572 (2000).
18. Chawengsaksophak, K., de Graaff, W., Rossant, J., Deschamps, J. & Beck, F. Cdx2 is essential for axial elongation in mouse development. *Proc. Natl. Acad. Sci. USA* **101**, 7641–7645 (2004).
19. Dufort, D., Schwartz, L., Harpal, K. & Rossant, J. The transcription factor HNF3 β is required in visceral endoderm for normal primitive streak morphogenesis. *Development* **125**, 3015–3025 (1998).
20. Kinkel, M. D., Eames, S. C., Alonzo, M. R. & Prince, V. E. Cdx4 is required in the endoderm to localize the pancreas and limit β -cell number. *Development* **135**, 919–929 (2008).
21. Martinez Barbera, J. P. et al. The homeobox gene Hex is required in definitive endodermal tissues for normal forebrain, liver and thyroid formation. *Development* **127**, 2433–2445 (2000).
22. Zaret, K. S. Genetic programming of liver and pancreas progenitors: lessons for stem-cell differentiation. *Nat. Rev. Genet.* **9**, 329–340 (2008).
23. Li, L. C. et al. Single-cell transcriptomic analyses reveal distinct dorsal/ventral pancreatic programs. *EMBO Rep.* **19**, e46148 (2018).
24. Jennings, R. E. et al. Laser capture and deep sequencing reveals the transcriptomic programmes regulating the onset of pancreas and liver differentiation in human embryos. *Stem Cell Rep.* **9**, 1387–1394 (2017).
25. Rodriguez-Seguel, E. et al. Mutually exclusive signaling signatures define the hepatic and pancreatic progenitor cell lineage divergence. *Genes Dev.* **27**, 1932–1946 (2013).
26. Tremblay, K. D. & Zaret, K. S. Distinct populations of endoderm cells converge to generate the embryonic liver bud and ventral foregut tissues. *Dev. Biol.* **280**, 87–99 (2005).
27. Angelo, J. R., Guerrero-Zayas, M. I. & Tremblay, K. D. A fate map of the murine pancreas buds reveals a multipotent ventral foregut organ progenitor. *PLoS One* **7**, e40707 (2012).
28. Chung, W. S., Shin, C. H. & Stainier, D. Y. Bmp2 signaling regulates the hepatic versus pancreatic fate decision. *Dev. Cell* **15**, 738–748 (2008).
29. 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).
30. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
31. Nowotschin, S. et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* **569**, 361–367 (2019).
32. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
33. Wang, X., He, Y., Zhang, Q., Ren, X. & Zhang, Z. Direct comparative analysis of 10X Genomics chromium and Smart-seq2. *bioRxiv* <https://doi.org/10.1101/615013> (2019).
34. Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
35. Li, L. et al. Single-cell RNA-seq analysis maps development of human germline cells and gonadal niche interactions. *Cell Stem Cell* **20**, 858–873 (2017).
36. 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).
37. Sherwood, R. I. et al. Prospective isolation and global gene expression analysis of definitive and visceral endoderm. *Dev. Biol.* **304**, 541–555 (2007).
38. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).
39. Li, H. et al. Classifying *Drosophila* olfactory projection neuron subtypes by single-cell RNA sequencing. *Cell* **171**, 1206–1220 (2017).
40. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
41. Kanton, S. et al. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019).
42. Hebrok, M., Kim, S. K. & Melton, D. A. Notochord repression of endodermal Sonic hedgehog permits pancreas development. *Genes Dev.* **12**, 1705–1713 (1998).
43. Miki, R. et al. Fate maps of ventral and dorsal pancreatic progenitor cells in early somite stage mouse embryos. *Mech. Dev.* **128**, 597–609 (2012).
44. Hastie, T. & Stuetzle, W. Principal curves. *J. Am. Stat. Assoc.* **84**, 502–516 (1989).
45. Stafford, D. & Prince, V. E. Retinoic acid signaling is required for a critical early step in zebrafish pancreatic development. *Curr. Biol.* **12**, 1215–1220 (2002).
46. Zorn, A. M. StemBook (IOS Press, Amsterdam, 2008).
47. Jung, J., Zheng, M., Goldfarb, M. & Zaret, K. S. Initiation of mammalian liver development from endoderm by fibroblast growth factors. *Science* **284**, 1998–2003 (1999).
48. Calmont, A. et al. An FGF response pathway that mediates hepatic gene induction in embryonic endoderm cells. *Dev. Cell* **11**, 339–348 (2006).
49. Zhang, N. et al. Biodegradable polymeric micelles encapsulated JK184 suppress tumor growth through inhibiting Hedgehog signaling pathway. *Nanoscale* **7**, 2609–2624 (2015).
50. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
51. Wang, D. et al. Gene expression profile analysis in epilepsy by using the partial least squares method. *Sci. World J.* **2014**, 731091 (2014).
52. Franklin, V. et al. Regionalisation of the endoderm progenitors and morphogenesis of the gut portals of the mouse embryo. *Mech. Dev.* **125**, 587–600 (2008).
53. Pan, F. C. & Wright, C. Pancreas organogenesis: from bud to plexus to gland. *Dev. Dyn.* **240**, 530–565 (2011).
54. McDole, K. et al. In toto imaging and reconstruction of post-implantation mouse development at the single-cell level. *Cell* **175**, 859–876 (2018).
55. Matsushita, S. Fate mapping study of the endoderm of the 1.5-day-old chick embryo. *Roux's Arch. Dev. Biol.* **205**, 225–231 (1996).
56. Gu, G. et al. Global expression analysis of gene regulatory pathways during endocrine pancreatic development. *Development* **131**, 165–179 (2004).
57. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
58. de Bakker, B. S. et al. An interactive three-dimensional digital atlas and quantitative database of human development. *Science* **354**, aag0053 (2016).
59. Grapin-Botton, A. StemBook (IOS Press, Amsterdam, 2008).
60. Kimura, S. et al. The T/ebp null mouse: thyroid-specific enhancer-binding protein is essential for the organogenesis of the thyroid, lung, ventral forebrain, and pituitary. *Genes Dev.* **10**, 60–69 (1996).
61. Spence, J. R. et al. Sox17 regulates organ lineage segregation of ventral foregut progenitor cells. *Dev. Cell* **17**, 62–74 (2009).
62. Kirby, M. L. et al. Hensen's node gives rise to the ventral midline of the foregut: implications for organizing head and heart development. *Dev. Biol.* **253**, 175–188 (2003).
63. Piette, D., Hendrickx, M., Willems, E., Kemp, C. R. & Leyns, L. An optimized procedure for whole-mount in situ hybridization on mouse embryos and embryoid bodies. *Nat. Protoc.* **3**, 1194–1201 (2008).
64. Gross-Thebing, T., Paksa, A. & Raz, E. Simultaneous high-resolution detection of multiple transcripts combined with localization of proteins in whole-mount embryos. *BMC Biol.* **12**, 55 (2014).
65. de Soysa, T. Y. et al. Single-cell analysis of cardiogenesis reveals basis for organ-level developmental defects. *Nature* **572**, 120–124 (2019).
66. Takahashi, M., Makino, S., Kikkawa, T. & Osumi, N. Preparation of rat serum suitable for mammalian whole embryo culture. *J. Vis. Exp.* **90**, e51969 (2014).
67. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
68. Anders, S., Pyl, P. T. & Huber, W. HTSeq — a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
69. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
70. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
71. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
72. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
73. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
74. Croft, D. et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
75. Akima, H. Algorithm 761: scattered-data surface fitting that has the accuracy of a cubic polynomial. *ACM Trans. Math. Softw.* **22**, 362–371 (1996).