# Neighbor-specific gene expression revealed from physically interacting cells during mouse embryonic development

Junil Kim[1,2,3*], Michaela Mrugala Rothová[2*], Linbu Liao[1,2], Siyeon Rhee[4], Guangzheng Weng[1,5], Eyal David[6], Ido Amit[6], Morteza Chalabi Hajkarim[1,2], Joshua M. Brickman[2**] and Kyoung Jae Won[1,2**]

[1]Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark, [2]Novo Nordisk Foundation Center for Stem Cell Biology, DanStem, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdams vej 3B, 2200 Copenhagen N, Denmark, [3]School of Systems Biomedical Science, Soongsil University, 369 Sangdo-Ro, Dongjak-Gu, Seoul, Republic of Korea, [4]Department of Biology, Stanford University, Stanford, CA 94305, USA, [5]Department of Biology, the Bioinformatics Centre, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark, [6]Department of Immunology, Weizmann Institute of Science, Rehovot, Israel

* Equally contributed to this work

** To whom correspondence should be addressed to KJW (Tel: +45-3533-1419 ; Email: kyoung.won@bric.ku.dk) and JMB (Email: joshua.brickman@sund.ku.dk)

# ABSTRACT

Cells continuously communicate with the neighboring cells during development. Direct interaction of different cell types can induce molecular signals dictating lineage specification and cell fate decisions. The current single cell RNAseq (scRNAseq) technology cannot study cell contact dependent gene expression due to the loss of spatial information. To overcome this issue and determine cell contact specific gene expression during embryogenesis, we performed RNA sequencing of physically interacting cells (PICseq) and assessed alongside our single cell transcriptomes (scRNAseq) derived from developing mouse embryos between embryonic day (E) 7.5 and E9.5. Analysis of PICseq data identifies an interesting suite of gene expression signatures depending on neighboring cell types. For instance, neural progenitor (NP) cells expressed *Nkx2-1* when interacting with definitive endoderm (DE) and DE cells expressed *Gsc* when interacting with NP. Based on the identified cell contact specific genes, we devised a means to predict the neighboring cell types from individual cell transcriptome. We further developed spatial-tSNE to show the pseudo-spatial distribution of cells in a 2-dimensional space. In sum, we suggest an approach to study contact specific gene regulation during embryogenesis.

**Keyword**: single-cell RNA sequencing, PICseq, contact-specific expression, mouse embryonic development, cell-cell contact, spatial-tSNE, neighboring cell types

# INTRODUCTION

Cell-cell contact is important for cell-fate specification in development(*1–9*). Cell-cell communication including direct cell-cell contact orchestrates embryonic patterning, cell type specification and organ formation. Removing specific tissue or placing ectopic explants into various embryonic regions can change the fate of cells adjacent to the manipulated area. For instance, manipulation with the crucial signaling center, anterior visceral endoderm (AVE), causes defects in specification of forebrain identity later in development(*10*). It is long known that ectopic grafts of signaling centers such as the node or Spemann organizer can induce a secondary neural axis(*11*, *12*).

Even with the importance of various cell type interactions for development, it is still not easy to study gene expression changes in association with neighboring tissue. While the grafting experiments have significantly improved our understanding of the inductive ability of various neighboring cell types(*12*), these experiments require very sophisticated experimental setup and/or are not always feasible (e.g. human embryonic development).

Single cell transcriptional profiling has been successfully applied to identify cell types and their developmental trajectory during mouse organogenesis(*13–15*). However, the loss of spatial information in scRNAseq makes it difficult to trace the genes induced by interaction of different cell types. Co-expression of ligand-receptors pairs from scRNAseq data has been used to predict interacting cell types(*16*, *17*). However, there are diverse ways of cell communication besides ligand-receptor such as direct cell communication through gap junctions(*18*). It is still challenging to detect neighboring tissue specific gene expression changes.

Recently, RNA sequencing of multi-cells has been used to measure the transcriptome of physically interacting cells (PICs) without relying on ligand-receptor co-expression(*19–22*). ProximID identified the interacting cell types in the bone marrow and the small intestine in mice using the transcriptome of two or more PICs by applying mild dissociation of cells(*19*). PICseq was applied to

interrogate interactions of immune and epithelial cells in neonatal murine lungs(*22*) and to understand transcriptomic changes of liver endothelial cells (LECs) across liver zones(*20*). Cell interaction by multiplet sequencing (CIM-seq) was applied to identify interacting cell types in gut epithelium, lung and spleen(*23*).

While previous approaches using multi-cells have mainly focused on interacting cell types, we hypothesize that interacting multi-cells will retain gene expression derived from the physical interaction defined by these cell types. To identify cell-type contact dependent transcriptional profiles focusing on endoderm development, we dissected developing mouse embryos (at E7.5, E8.5 and E9.5), sorted FOXA2$^{Venus}$ expressing cells and obtained PICseq. PICseq captured both homotypic and heterotypic cell clumps physically interacting with each other. Interestingly, we identified various sets of genes expressed in the heterotypic PICs but not each individual cell type. Computational analysis identified the cell types expressing unique gene sets specific to their neighbors. For instance, Lhx5 and Nkx2-1 were expressed exclusively in the NP cells that physically interact with DE cells and Gsc was expressed in the DE cells interacting with NP cells. Some of these neighboring cell type specific genes were associated with the development of specific embryonic regions. Validation using Geo-seq which provides transcriptome from the dissected developing mouse embryos(*24*) confirmed that we had successfully identified spatially organized sets of cells and neighboring cell specific genes. Notably, we were able to predict neighboring cell types from scRNAseq based on the cell-contact specific genes. We further imitated the anatomy of the mouse embryos by visualizing cell contact information and neighboring cell specific gene expression in a modified t-distributed stochastic neighbor embedding plot (spatial-tSNE). Our results suggest that the local environment information retained in the transcriptome of individual cells can be used to reconstruct spatial gene regulation during development.

# RESULTS

**Mapping of PICseq data onto cell types in the mouse embryo**

To identify transcriptional profiles influenced by cell contact during development focusing on endoderm, we exploited mouse line in which a Venus fluorescent protein had been fused to the endoderm regulator FOXA2 to identify FOXA2 expressing endoderm cells from embryos ranging from E7.5 - E9.5 (Fig. 1A). We obtained PICseq after mild dissociation of these mouse embryos followed by fluorescence-activated cell sorting (FACS) against FOXA2$^{Venus}$ and cell size to filter out potential single cells.

After quality control, RNA sequencing data from 367 PICs were obtained using massively parallel RNA single-cell sequencing (MARS-seq) technology(*25, 26*). To analyze the PICseq data, we used the scRNAseq (total 6,288 cells) from the mouse embryo obtained at matching time points (manuscript submitted). From the scRNAseq, we identified 13 clusters: neural progenitors (NP), foregut (FG), midgut (MG), hindgut (HG), definitive endoderm (DE), liver (L), extra-embryonic visceral endoderm (ExVE), embryonic visceral endoderm (EmVE), primitive streak (PS), node (N), notochord (NC), parietal endoderm (PE) and one undefined cluster (Fig. 1B, C and Supplementary Fig. 1). From the cluster analysis, we identified 545 cell-type specific marker genes (FDR < 0.01 & log ratio > 0.4) (Supplementary Table S1).

To annotate the cell types of the PICs, we used the cell type marker genes identified from scRNAseq clustering (Fig. 1C). For this, multiclass support vector machines (SVMs) were trained with artificial doublets from two randomly selected single cells for the three time points. The trained SVMs classified the 367 PICs into homotypic and heterotypic combinations of cell types (Fig. 1D). The error rate of 10-fold cross-validation for the SVMs were 17.27%, 18.76%, and 28.56% in E7.5, E8.5, and E9.5, respectively, suggesting comparable performance with the previous deconvolution approach(*22*). The marker genes were well expressed in the annotated PICs (Fig. 1D). Especially, the frequencies of the heterotypic combinations identified using PICs were significantly high compared with the combination of erroneous doublets identified using DoubletFinder(*27*) (Supplementary Fig. 2). The

observed heterotypic combinations such as ExVE+EmVE (at E7.5) and NP+DE (at E8.5) well match the neighboring tissues in developing mouse embryos (Fig. 1D).

**PICseq enables detection of cell-cell contact specific gene expression**

To examine cell-cell contact specific gene expression, we investigated genes highly expressed in the heterotypic PICs but not in each individual cell type. For the heterotypic PICs at least 10 (ExVE+EmVE in E7.5, NP+DE in E8.5, NP+NC in E8.5, MG+HG in E9.5, and FG+MG in E9.5), we investigated neighboring cell dependent gene expression (Fig. 2 and Supplementary Fig. 3-7). Interestingly, we identified 167 genes highly expressed in the heterotypic PICs compared with the expected expression levels for individual cell types obtained from scRNAseq (FDR < 0.01 & log ratio > 0.5) (Supplementary Table S2). For instance, the heterotypic PICs for NP+DE expressed genes such as *Lhx5*, *Nkx2-1*, and *Gsc*. These PICs expressed the marker genes for NP and DE: *Crabp1* and *Col8a1* for NP, and *Trh* and *Slc2a3* for DE (Fig. 2A, 2B, and Supplementary Fig. 3). Interestingly, some of the identified neighboring cell type specific genes are known to play crucial roles during development. *Lhx5* is known to promote forebrain development(*28*). Nkx2-1 is required for the proper specification of interneuron subtypes(*29*). Finally, *Gsc* is a known marker of anterior endoderm(*30*) and pre-chordal plate(*31*) and has been implicated in ESC differentiation to definitive endoderm differentiation(*32*, *33*).

The 54 genes highly expressed in the heterogeneous PIC of NP+DE at E8.5 were linked to endodermal or neuroectodermal embryo development including gland development (*Sox3*, *Nkx2-1*, and *Wnt4*, p-value: 1.1E-3) and nervous system development (*Sox3*, *Gbx2*, *Dpysl5*, *Sall3*, *Ptn*, and *Apba2*, p-value: 1.38E-3) based on the gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment tests(*34*) (Fig 2D). The 39 NP+NC contact-specific genes were mainly associated with mesodermal development (*Mesp1*, *Meox1*, *Lef1*; Supplementary Fig. 6). Our results suggest that PIC-seq could be identifying contact specific genes induced during embryonic differentiation.

**Contact-specific gene expression is spatially localized at the boundary regions between two cell types**

To validate the spatial expression of contact-specific marker genes, we used Geo-seq(*24*) which contains transcriptome from 83 dissected regions in the developing mouse embryos at E7.5 (Supplementary Fig. 8-10). DE marker genes (*Cer1* and *Trh*) were strongly expressed in anterior endoderm (EA), while NP marker genes (*Nkx6-1*, and *Crabp1*) were strongly expressed interior to the endoderm in Geo-seq (Fig. 2C, Supplementary Fig. 8). Intriguingly, the contact specific genes (*Lhx5* and *Tubb3*) were expressed in between the regions expressing NP and DE specific genes (Fig. 2C, Supplementary Fig. 9). The averaged profiles for NP+DE genes also showed strongest expression of the contact specific genes in between NP and DE marked regions in the Geo-seq (Supplementary Fig. 10A).

In addition, we found that the ExVE+EmVE specific genes were mainly expressed in between the regions marked by the two cell types (upper part of the corn plot) in the Geo-seq dataset (Supplementary Fig. 10B). While the nature of the GEO-seq dataset is such that it does not include structures like gut or notochord, we identify co-expression of contact dependent genes, e.g. the NP+NC specific genes were highly expressed in the posterior mesoderm (Supplementary Fig. 10C), which was located between the posterior ectoderm and node region (bottom of the corn plot). Our results indicate that the contact specific genes obtained from PICseq reflect the actual spatial expression.

**Prediction of the neighboring cell type from scRNAseq**

Based on the contact specific gene expression, we designed an approach to predict neighboring cell types by using the transcriptome of single cells (see Method). Briefly, the contact specific genes in

the PICseq were used to train a multiclass SVM. The trained SVM predicts the interacting cell types of a cell by forming artificial doublets and voting from the SVM output (see Method).

Among the 767 NP cells, 85 were predicted to neighbor with DE, 123 interacted with NC and the majority (559) interacted with other NP cells (Fig. 3A). Among the 677 DE cells, 252 were predicted to interact with NP (Fig. 3B). Notably, neighboring cell type prediction further told us the cell type expressing contact specific genes. Among the contact specific genes in the heterotypic NP+DE PICs, *Lhx5*, *Pou3f1*, *Mest*, *Ptn*, *Nkx2-1*, and *Slc1a3* were from the NP cells and *Gsc*, *Sox3*, and *Rax* were from the DE cells (Fig. 3A-B and Supplementary Table S3). These analyses further predicted the list of neighboring cell specific genes expressed in each cell type. For instance, NP cells expressed *Lhx* and *Nkx2-1* when interacting with DE and *Prtg* and *Gas1* when interacting with NC cells (FDR<0.01) (Fig. 3A).

To validate these contact specific genes, we performed co-staining of FOXA2 (a DE marker), SOX2 (an NP marker), and NKX2-1 (a contact specific gene) in the mouse embryo at E8.25. FOXA2 was expressed in DE as well as the floor plate that connects to NP. SOX2 was expressed stronger in NP (compared with DE). NKX2-1 was expressed in the part of NP cells when they are proximal to DE, consistent with our prediction (Fig. 3C).

We also predicted the neighboring cell type for FG, MG, HG, NC, and EmVE cells using the same strategy (Supplementary Fig. 11). We further tested if the trained SVM can annotate the neighboring cell types for publicly available scRNAseq data from developing mouse embryos(*13*, *15*). After the annotation, we still found the distinctive groups of DE cells expressing contact specific genes when contacting NP such as *Rax* and *Gsc* in these independent datasets (Supplementary Figs. 12 and 13). Our results indicate that scRNAseq retains the information about the neighboring cell type even after cells are isolated. In sum, we observed diverse repertoire of contact specific genes depending on the neighboring cell types.

**Visualizing spatio-structure of tissue using spatial-tSNE**

Current 2-dimensional (2D) embeddings for scRNAseq cannot represent spatial information. To visualize neighboring cells and the neighboring cell type specific expression profiles, we developed spatial t-distributed stochastic neighboring embedding (spatial-tSNE) by modifying the similarity probability of cell pairs that are classified into neighboring cells. Compared with conventional visualization approaches based on the transcriptomic similarities (*e.g.* Fig. 1B), spatial-tSNE provides the information about interacting cell types in the mouse embryos. For instance, the spatial-tSNE showed the physical interactions between EmVE and DE (Fig. 4A), while they were distally located in the classical UMAP plot (Fig. 1B). Besides, the spatial-tSNE plot represented spatial layout of NP+DE, MG+NC and DE+EmVE, which is consistent with the anatomy of the mouse embryos.

The spatial-tSNE plot recapitulates the neighboring cell type specific expression patterns in 2D embedded dimensions (Fig. 4A). The average expressions of NP+DE, MH+HG, and NP+NC markers were localized near the border of the corresponding two cell types (Supplementary Fig. 14). Spatial-tSNE visualizes the neighboring cell-type specific factors. For instance, *Lhx5* is expressed in NP cells close to DE and *Gsc* is expressed more profoundly in DE cells close to NP (Fig. 4B).

# DISCUSSION

Cells are influenced by the neighboring cells in many ways including cell size, stiffness, and mechanical forces(*35*, *36*). During embryogenesis, cells have to coordinate with each other to express genes in the correct spatial and temporal context. During development cells can influence each other's gene expression by exploiting various signaling molecules, direct cell-cell contact(*37*) or reconfiguring mechanical environment(*38*, *39*).

We questioned if there are genes expressed when cells contact with each other. To obtain cell-cell contact information, we used PICseq and developed computational algorithms to identify

neighboring cell type specific gene expression. PICseq, by retaining cell contact, enabled us to predict the neighboring cell type specific gene expression. Our predictions indicate that cells present specific gene expression patterns depending on the neighboring cell type. (Fig. 3A). For instance, NP cells expressed *Lhx5* and *Nkx2-1* when neighboring with DE and *Gas1* when neighboring with more posterior NC.

To validate our observation in an unbiased way, we used Geo-seq(*24*). The neighbor-specific genes we identified were spatially expressed between the regions marked for each cell type (Fig. 2). Even though Geo-seq(*24*) is limited to the mouse embryos at E7.5, it was sufficient to highlight the likely spatial location of pairs of interactors for NP+DE, ExVE+EmVE and NP+NC (Supplementary Fig. 10).

Interestingly, among the predicted neighboring cell type specific genes, we identified several factors known to play important roles during development. For instance, *Lhx5*, a gene expressed in NP cells when contacting DE, is known to promote forebrain development by regulating the Wnt signaling pathway(*28*). Knockdown of *Lhx5* resulted in apoptotic hypothalamic development(*40*). *Nkx2-1*, another gene expressed in NP cells when contacting DE, is also known for its role in response to dorsal-ventral patterning in the neural tube and specifying cortical interneuron subtypes(*29*). Also, DE cells interacting with NP cells expressed *Gsc* (Fig. 2A and 3B) which has a role in anterior endoderm(*30*), pre-chordal plate(*31*), definitive endoderm differentiation(*32*) and foregut formation(*33*). Our results may indicate that cell contact can trigger cell type specification during embryogenesis.

Our study indicates that a cell encodes details about its local environment in its transcriptome. For instance, single NP cells identified as interacting with DE retained the expression of *Nkx2-1* even after cell isolation (Fig. 2A and 3A). Based on the local information detailed in the transcriptome of a cell, we were able to predict its neighboring cell type (Fig. 3A, B). The trained model was further applied to identifying neighboring cell types for publicly available scRNAseq during embryogenesis(*13*,

*15*) (Supplementary Figs. 12 and 13). This indicates that the contact specific genes identified from PICseq can be used as the reference to re-annotate the neighboring cell types of public scRNAseq.

Gene expression varied between cells based on their neighbors. In our study, NP cells interacting with DE expressed *Lhx5* and *Nkx2-1*, and those interacting with NC expressed *Prtg* (Fig. 3A). These findings may underlie anterior-posterior axis inducing activities of these derivatives of the node. The anterior endoderm emerges very early in gastrulation to pattern the presumptive anterior neural plate, while the NC emerges later, patterning more posterior locations along the neural tube. Deconstructing positional information within the transcriptome could provide a detailed map of cells localized to axis promoting organizing centers that emerge in embryonic development. Given the role of anterior endoderm in patterning the nascent neural plate(*10*) and NC for the patterning of neural tube(*41*), our identification of NP+DE and NP+NC supports the inductive profile of these cells.

To confirm weather our method can be generalized to other systems, we applied this approach on a published PIC-seq dataset of hepatocytes (HC) and liver endothelial cells (EN)(*20*). We identified 69 HC+EN contact-specific genes, which are associated with cholesterol metabolism (*Cyp27a1*, *Apoh*, *Apoc1*, *Angptl3*, and *Apoe*, p-value: 7.92E-7; Supplementary Fig. 15 and Supplementary Table S4). *Angptl3* a liver-specific secreted protein is known to induce endothelial cell migration and adhesion(*42*), which was not identified in the original paper. To validate this finding, we performed staining analysis using ANGPTL3 and PECAM1 and endothelial cell marker. We found stronger signals for ANGPTL3 at the regions where blood vessels are located (marked by PECAM1), validating that *Angptl3* is highly expressed in the hepatocytes interacting with endothelial cells (Supplementary Fig. 15D). This shows that our strategy can be used for other types of data.

A widely used approach to understand cell-communication is to use ligand-receptor pairs(*16*, *17*). We did not see clear ligand-receptor pairs in the contacting cell pairs in our data. It is of note that PICseq enabled investigation of cell contact dependent gene expression which cannot be studied using ligand-receptor pairs. The use of co-expression of ligand-receptor pairs only shows the most frequently

interacting cells and does not explain specific cells engaged in cell communication. The cell contact information by PICseq may provide another layer of information about cell-communication.

We developed spatial-tSNE to visualizes the spatial proximity that we predicted. Previous visualization methods locate cells mainly based on transcriptomic similarities. The UMAP plot using scRNAseq data well represented spatial organization of EmVE and ExVE as well as FG, MG, and HG, because their gene expression is related to their spatial interactions (Fig. 1B). However, transcriptome-based visualization could not represent the physical interaction of NP+DE, MG+NC and NP+EmVE, while these are well presented in the spatial-tSNE (Fig 4A). Consistent with this representation, the contact specific genes are found in association with their locations in the spatial-tSNE plot (Figure 4B). Spatial-tSNE depends on the prediction of neighboring cell types (Fig. 3A and B, Supplementary Fig. 11). In this work, we used 367 PICs for training. The performance of neighboring cell type prediction can be further improved as we accumulate more PICs for training. Even though spatial-tSNE cannot represent the real 3D structure of the tissue, it provides a more comprehensive map for context dependent relationships inherent in mammalian development.

The history of developmental biology is based on a large number of embryonic grafting experiments to define inductive interactions that occur during development(*12*). Grafting experiments were pivotal in our understanding of how signals produced from one cell can be illicit inductive or provides patterning responses in another cell. However, grafting experiments are technically challenging and inherently limited to a pre-defined set of interactions. Here we take an unbiased approach to understand developmental context, producing spatial-tSNEs to provide an unbiased catalogue of potential developmental interactions. Through assessing these interactions, one could develop a comprehensive map of embryonic induction providing a set of all possible sites. While directionality can only be inferred by experiment, we present an unbiased approach to study spatial gene regulation during development.

# METHODS

## Multiplet cell isolation

Mouse FOXA2$^{Venus}$ embryos were collected between embryonic days (E) 7.5 and 9.5. The E9.5 embryos were dissected in order to enrich the sample with gut endoderm cells. Embryos were dissociated with Accutase (Sigma) into multiplet cells immediately after collection. The collected embryos were mixed with mouse embryonic stem cells (ESC), which were counterstained with CellVue Maroon Cell Labeling Kit (Thermofisher, # 88-0870-16) to increase the number of cells in the sample in order to avoid loss of the scarce FOXA2$^{Venus}$ positive cells by spinning. The samples were then incubated with pre-warmed Accutase at 37C. For the multiplet cell dissociation, the samples were incubated for 4 mins with 1ml of Accutase and pipetted up and down carefully to ensure cells were not dissociated into single cells. The Accutase was diluted by adding of 3ml of FACS buffer (10% FBS in PBS) and span down. The cells were washed with FACS buffer two more times and transferred to a FACS tube with FACS-DAPI buffer for the FACS process.

## Flow cytometry and multiplet cell capturing

Multiplet cells were isolated from FOXA2$^{Venus}$ mouse embryos. Cells were sorted using a BD FACS Aria III. FSC/SSC gates were used to define a homogeneous population, FSC-H/FSC-W gates were used to include multiplets, remove singlets and dead cells were excluded based on DAPI inclusion. Sorting speed was kept at 100-300 events/s to eliminate sorting two or more drops containing cells into one well. Single drop deposition into the 384 well plates was verified colorimetrically based on a previously published protocol(*43*). Cells were sorted directly into lysis buffer containing first RT primer and RNase inhibitor, immediately frozen and later processed by MARS-seq1 protocol as described previously(*25*).

**Data processing of scRNAseq and PICseq data of mouse embryo**

All scRNAseq libraries were sequenced using Illumina NextSeq 500. Sequences were mapped to mouse mm9 genome, de-multiplexed and filtered as previously described(*25, 26*). We estimated the level of spurious unique molecular identifier (UMIs) in the data using statistics on empty MARS-seq wells as previously described(*25*). Mapping of reads was done using HISAT (version 0.1.6)(*44*); reads with multiple mapping positions were excluded.

Among the transcriptome of 6,600 single cells and 382 PICs, we filtered the low-quality samples that had UMI counts over $2^{17}$ or less than 256 $2^8$. The remaining samples contained 6,288 single cells and 367 PICs. To remove any batch effect, we used the Seurat v3 standard integration workflow(*45*). The 13 clusters were obtained using a graph-based Louvain clustering algorithm.

**SVMs for classification**

For the classification of PICs, we trained a multiclass classifier for SVMs using a MATLAB (version R2020a) function 'fitcecoc'. The classifier consists of multiple SVM binary learners in a one-vs-one design. We trained three SVMs for each stage (E7.5, E8.5, and E9.5) of PICseq and scRNAseq data. We used the major clusters for each stage (DE, N, ExVE, EmVE, PS, PE for E7.5; NP, FG, DE, MG, HG, NC for E8.5; FG, MG, HG, L, NC for E9.5) and the top 5 DEGs for each cluster. We run 10-fold cross-validation for the trained models using a MATLAB function 'crossval' and calculated the error rate using 'kFoldLoss' with 10. We calculated significance of the frequencies of PICs against doublets identified from scRNAseq using Fisher's exact test.

**SVMs for neighboring cell type prediction**

For the prediction of the identity of neighboring cells, we applied multiclass SVM using 'fitcecoc' function in MATLAB R2020a. We only used the contact-specific genes for training and prediction. The SVMs were trained using the PICseq data with their annotations. To predict the neighboring cell

type from the single cell transcriptome, we made artificial PICseq by mixing the transcriptome of the cell of interest with all other cells one by one. The voting schemed (i.e., most frequent) is used to assign the neighboring cell type.

**Statistical analysis and enrichment analysis**

The p-value for cell-type specific contact-specific marker genes was calculated by using a two-sided Wilcoxon rank sum test. We used Enrichr(*34*), an enrichment analysis tool, to investigate the enriched GO terms and KEGG pathways for each marker gene group.

**Staining of mouse embryo**

E8.5 embryos were isolated in PBS and then fix overnight in 4% paraformaldehyde (PFA) overnight at 4 °C. The following day, embryos were washed in PBT (PBS containing 0.1% Tween-20), dehydrated in an ascending methanol sequence, xylene treated, embedded in paraffin, and sectioned at 6.5 µm. Immunofluorescence was performed on 6.5 µm deparaffinized sections. Sections were subjected to antigen retrieval in Tris buffer pH 10.2 for 10 min, washed in 0.1% PBT and incubated in blocking buffer (0.5% milk powder, 99.5% PBT) for 2 h at room temperature. Primary antibodies were incubated in blocking buffer overnight at 4 °C. The following day, the sections were washed three times with PBT and incubated for 1 h with corresponding secondary antibodies in blocking buffer at room temperature. After three washes in PBT, DAPI (Sigma-Aldrich, 1:2000) was added to counter-stain the nuclei. The sections were mounted using Prolong Gold Antifade Reagent (Invitrogen, P36934) and imaged using Zeiss LSM-700 confocal microscope. The following primary antibodies were used: FOXA2 (Thermofisher, 7H4B7, 1:250), SOX2 (eBioscience, 14-9811-80, 1:250), NKX2-1 (Abcam, ab76013, 1:250). Secondary antibodies were Alexa Fluor conjugates 488, 555, and 647 (Life Technologies) at 1:500.

**Staining of mouse liver tissue**

Mouse liver was snap-frozen by embedding in OCT Tissue Tek (Sakura Finetechnical, Torrance, CA) after 25% sucrose treatment using isobutane chilled in dry ice. 7-μm sections were cut by Leica cryostat (Leica, Bannockburn, IL). Specimens were fixed with ice-cold acetone, followed by blocking and permeabilization with 2% BSA and 1% Triton X-100 in PBS, and incubated with primary antibodies against Angptl3 (1:100; Santa Cruz, CA), and PECAM (1:500; R&D system, MN). The

target proteins were visualized with the secondary antibodies conjugated with fluorescence (Alexa fluor 488 and 594, 1:400; Invitrogen, CA) and Hoechst nuclear stain. Fluorescence images were taken and processed using a fluorescent microscope (Olympus, CA).


**spatial-tSNE**

Spatial-tSNE is designed for visualization of neighboring cells in scRNAseq data, which is modified from tSNE(*46*) by considering neighboring cell information. In the original tSNE, the 2D embeddings is obtained by rearranging each cell based on the probability of the pairwise similarities of the cells on the original high-dimensional space. The pairwise similarity of two cells $i$ and $j$ in the high-dimensional space is defined by the probability, $p_{ij} = \frac{exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k \neq i} exp\left(-\|x_i - x_k\|^2/2\sigma_i^2\right)}$ where $x_i$ is the data points of cell $i$ and $\sigma_i$ is the variance of the Gaussian which is centered on data point $x_i$. Spatial-tSNE shows the clustering and spatial information at the same time by changing the similarity probability for the cells that are predicted to neighbor with each other. For this, we changed the similarity probability of two neighboring cells to the maximum probabilities of all pairs. With these modified probabilities, the predicted neighbor cells can be located at the border of the two clusters without ruining the relative position of other cells. To clearly visualize neighbored cells, spatial-tSNE aligns the neighboring cell pairs along a line at the border of the two clusters.

# AVAILABILITY

Source code and input files for the PIC classification and the neighboring cell type prediction are available at https://github.com/neocaleb/NicheSVM. Source code for spatial-tSNE is available at https://github.com/wgzgithub/sp_tSNE. The PIC-seq data can be downloaded from Gene Expression Omnibus database with accession number (GSE182393).

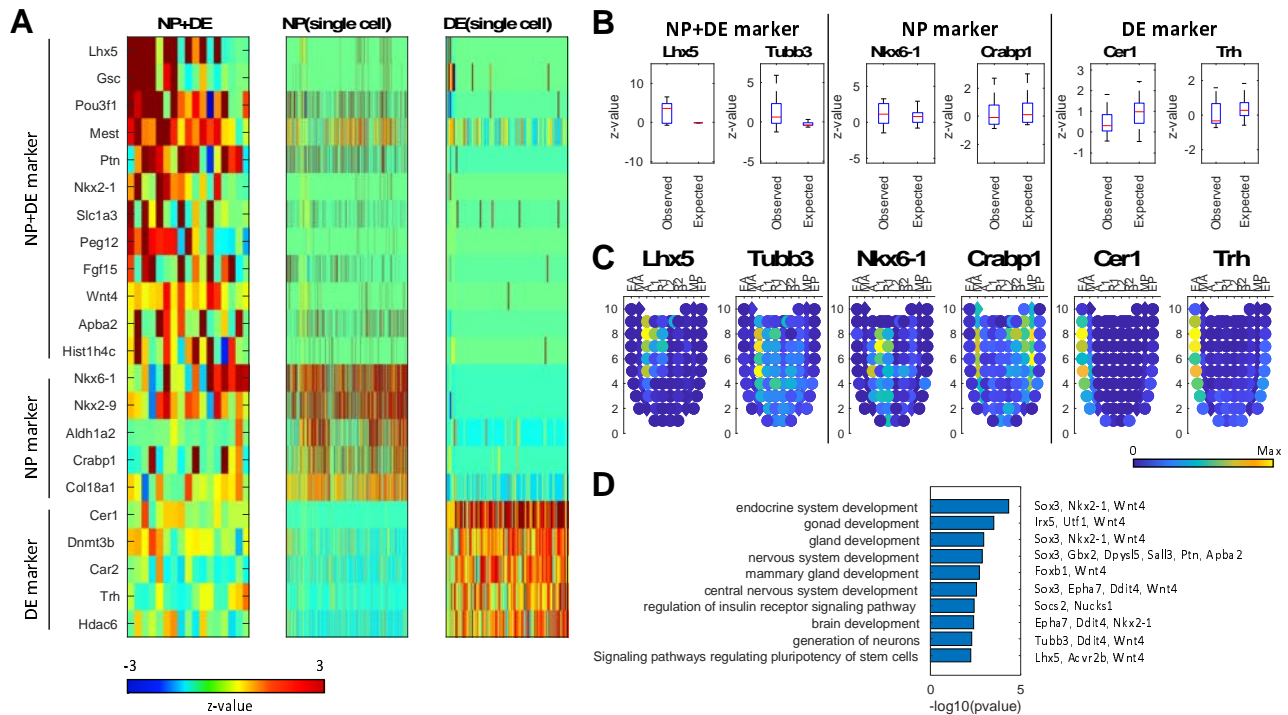# ACKNOWLEDGEMENTS
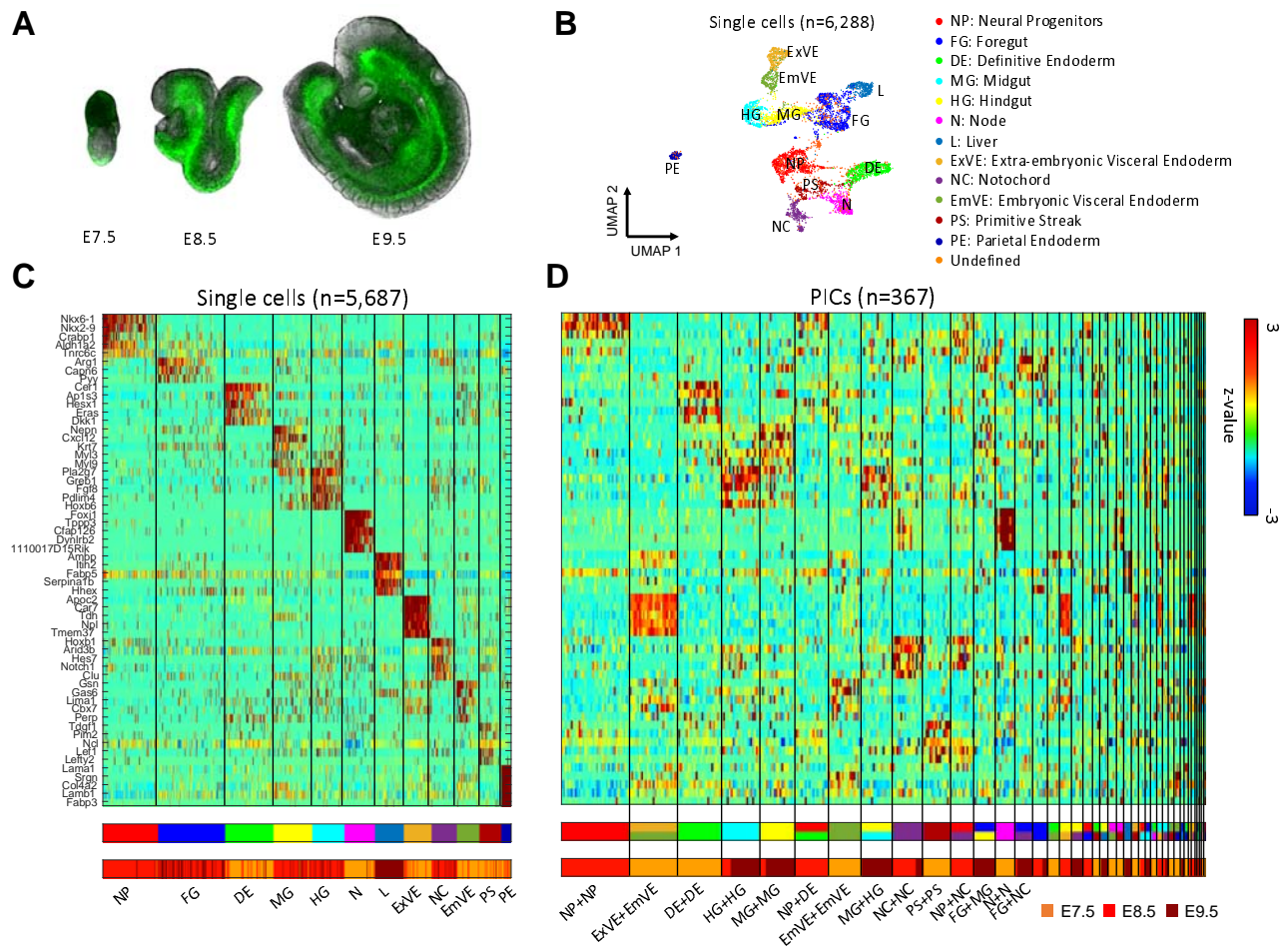
# FIGURES



**Figure 1. PICseq for developing mouse embryos.** (**A**) A cartoon of mouse embryos (E7.5, E8.5 and E9.5) with fluorescently tagged FOXA2$^{\text{Venus}}$. (**B**) A clustering result using scRNAseq data from the mouse embryos. (**C**) A heatmap of the top marker genes for the identified cell types (**D**) The gene expression and the annotation for the PICseq data.

**Figure 2. PICseq analysis identified genes highly expressed in NP+DE.** (**A**) PICseq of NP+DE showed contact specific expression as well as the marker genes for NP and DE. (**B**) contact specific expression levels are significantly high for the PICseq compared with the expected expression levels obtained from scRNAseq. (**C**) Validation of the contact-specific marker genes using Geo-seq data. EA: Anterior Endoderm; MA: Anterior Mesoderm; A: Anterior; L1: Anterior Left Lateral; R1: Anterior Right Lateral; L2: Posterior Left Lateral; R2: Posterior Right Lateral; P: Posterior; MP: Posterior Mesoderm, EP: Posterior Endoderm. (**D**) GO and KEGG pathway terms enriched in the contact-specific genes.
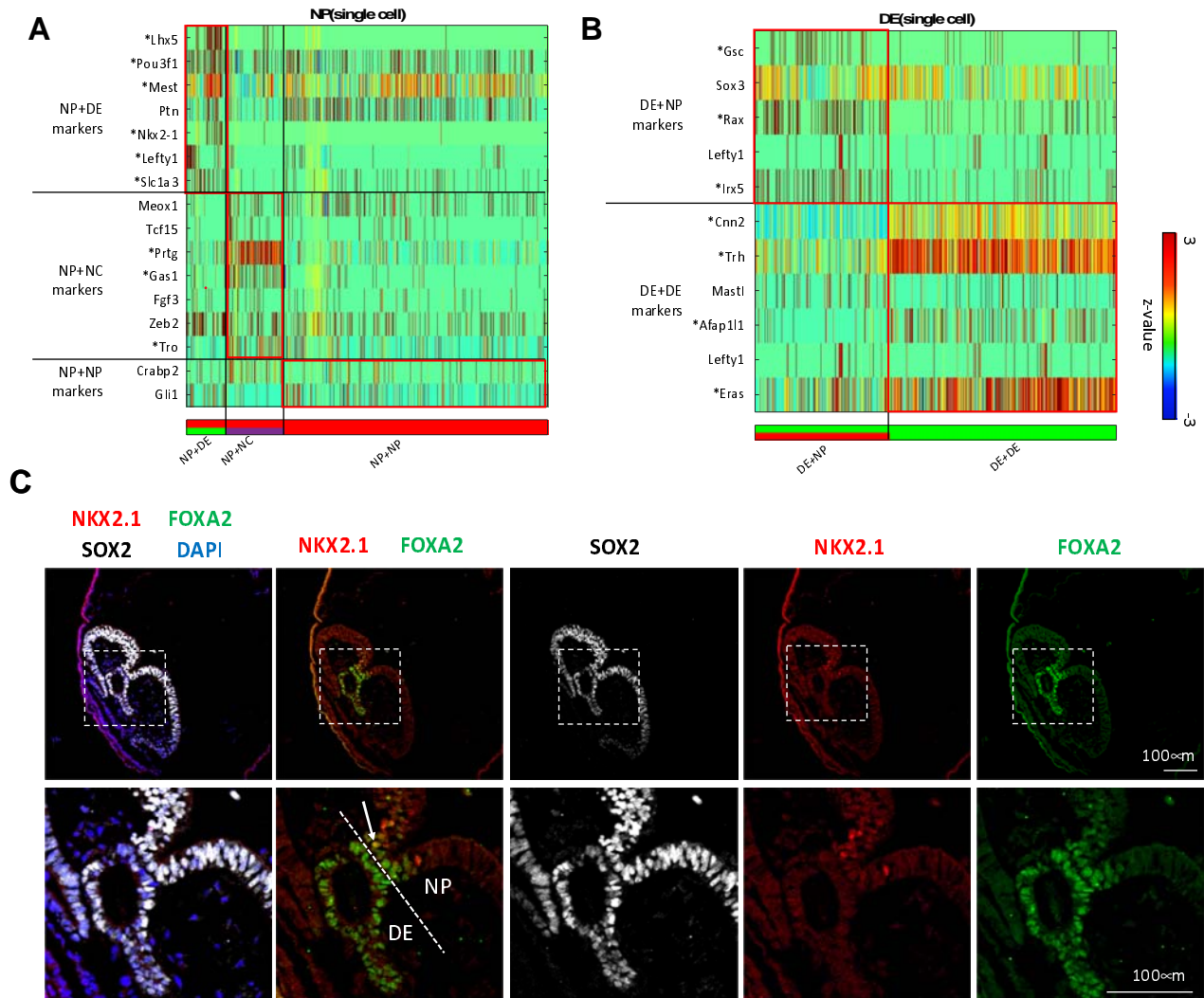
**Figure 3. Neighboring cell type prediction from scRNAseq** (**A**) The predicted neighboring cell types of NP cells (**B**) The predicted neighboring cell types of DE cells (**C**) Nkx2-1 is expressed in the NP (SOX2+) cells contacting with DE (FOXA2+).
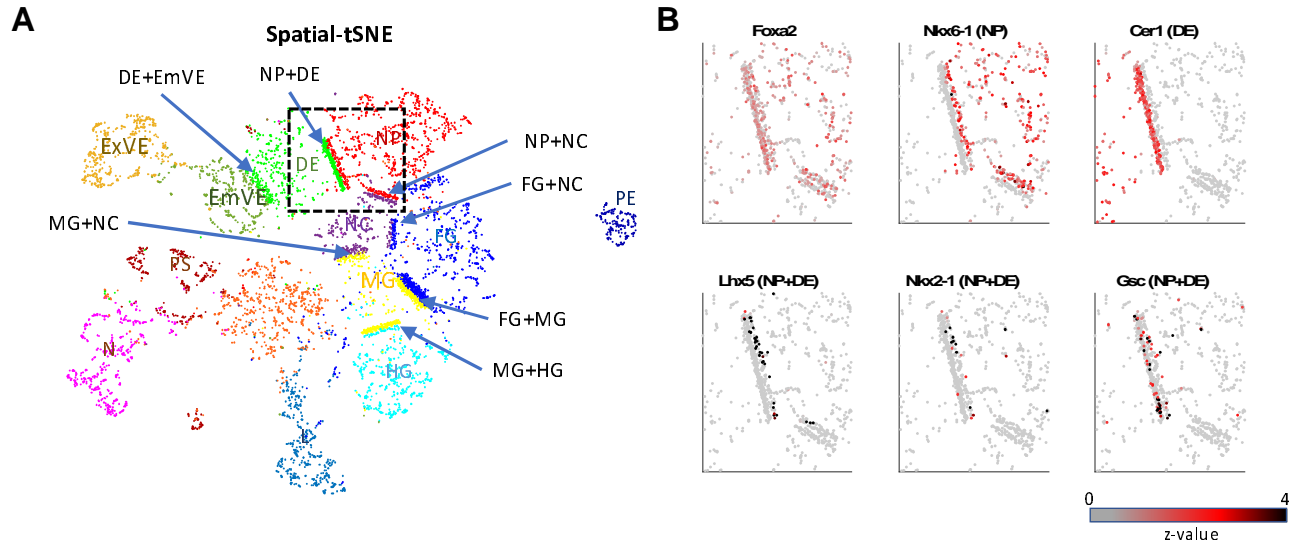
**Figure 4. Visualizing spatio-structure of tissue using spatial-tSNE for scRNAseq** (**A**) A spatial-tSNE plot recapitulating the spatial distribution of cells in mouse embryos. (**B**) Expression patterns of NP, DE, and NP+DE markers on the spatial-tSNE plot.

# REFERENCES

1.    S. G. Babb, J. A. Marrs, E-cadherin regulates cell movements and tissue formation in early zebrafish embryos. *Dev. Dyn.* **230**, 263–277 (2004).

2.    B. Goldstein, Induction of gut in Caenorhabditis elegans embryos. *Nature*. **357**, 255–257 (1992).

3.    N. Guisoni, R. Martinez-Corral, J. Garcia-Ojalvo, J. de Navascués, Diversity of fate outcomes in cell pairs under lateral inhibition. *Dev.* **144**, 1177–1186 (2017).

4.    R. K. L. Larue, C. Antos, S. Butz, O. Huber, V. Delmas, M. Dominis, A role for cadherins in tissue formation | Development. *Development*. **122**, 3185–3194 (1996).

5.    M. Michel, I. Raabe, A. P. Kupinski, R. Pérez-Palencia, C. Bökel, Local BMP receptor activation at adherens junctions in the Drosophila germline stem cell niche. *Nat. Commun.* **2**, 1–12 (2011).

6.  O. Shaya, U. Binshtok, M. Hersch, D. Rivkin, S. Weinreb, L. Amir-Zilberstein, B. Khamaisi, O. Oppenheim, R. A. Desai, R. J. Goodyear, G. P. Richardson, C. S. Chen, D. Sprinzak, Cell-Cell Contact Area Affects Notch Signaling and Notch-Dependent Patterning. *Dev. Cell*. **40**, 505-511.e6 (2017).

7.  R. O. Stephenson, Y. Yamanaka, J. Rossant, Disorganized epithelial polarity and excess trophectoderm cell fate in preimplantation embryos lacking E-cadherin. *Development*. **137**, 3383–3391 (2010).

8.  O. Tassy, F. Daian, C. Hudson, V. Bertrand, P. Lemaire, A quantitative approach to the study of cell shapes and interactions during early chordate embryogenesis. *Curr. Biol.* **16**, 345–358 (2006).

9.  M. D. Yoder, B. M. Gumbiner, Axial protocadherin (AXPC) regulates cell fate during notochordal morphogenesis. *Dev. Dyn.* **240**, 2495–2504 (2011).

10. P. Thomas, R. Beddington, Anterior primitive endoderm may be responsible for patterning the anterior neural plate in the mouse embryo. *Curr. Biol.* **6**, 1487–1496 (1996).

11. H. Spemann, H. Mangold, Induction of embryonic primordia by implantation of organizers from a different species. *Int. J. Dev. Biol.* **45**, 13–38 (2001).

12. G. E. Solini, C. Dong, M. Saha, Embryonic transplantation experiments: Past, present, and future. *Trends Dev. Biol.* **10**, 13–30 (2017).

13. S. Nowotschin, M. Setty, Y. Y. Kuo, V. Liu, V. Garg, R. Sharma, C. S. Simon, N. Saiz, R. Gardner, S. C. Boutet, D. M. Church, P. A. Hoodless, A. K. Hadjantonakis, D. Pe'er, The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature*. **569**, 361–367 (2019).

14. J. Cao, M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers, C. Trapnell, J. Shendure, The single-cell transcriptional landscape of mammalian organogenesis. *Nature*. **566** (2019), doi:10.1038/s41586-019-0969-x.

15. B. Pijuan-Sala, J. A. Griffiths, C. Guibentif, T. W. Hiscock, W. Jawaid, F. J. Calero-Nieto, C. Mulas, X. Ibarra-Soria, R. C. V. Tyser, D. L. L. Ho, W. Reik, S. Srinivas, B. D. Simons, J. Nichols, J. C. Marioni, B. Göttgens, A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*. **566**, 490–495 (2019).

16. R. Vento-Tormo, M. Efremova, R. A. Botting, M. Y. Turco, M. Vento-Tormo, K. B. Meyer, J. E. Park, E. Stephenson, K. Polański, A. Goncalves, L. Gardner, S. Holmqvist, J. Henriksson, A. Zou, A. M. Sharkey, B. Millar, B. Innes, L. Wood, A. Wilbrey-Clark, R. P. Payne, M. A. Ivarsson, S. Lisgo, A. Filby, D. H. Rowitch, J. N. Bulmer, G. J. Wright, M. J. T. Stubbington, M. Haniffa, A. Moffett, S. A. Teichmann, Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*. **563**, 347–353 (2018).

17. M. Efremova, M. Vento-Tormo, S. A. Teichmann, R. Vento-Tormo, CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc*. **15**, 1484–1506 (2020).

18. J. C. Hervé, M. Derangeon, Gap-junction-mediated cell-to-cell communication. *Cell Tissue Res.* **352** (2013), pp. 21–31.

19. J. C. Boisset, J. Vivié, D. Grün, M. J. Muraro, A. Lyubimova, A. Van Oudenaarden, Mapping the physical network of cellular interactions. *Nat. Methods*. **15**, 547–553 (2018).

20. K. B. Halpern, R. Shenhav, H. Massalha, B. Toth, A. Egozi, E. E. Massasa, C. Medgalia, E. David, A. Giladi, A. E. Moor, Z. Porat, I. Amit, S. Itzkovitz, Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. *Nat. Biotechnol*. **36**, 962–970 (2018).

21. B. M. Szczerba, F. Castro-Giner, M. Vetter, I. Krol, S. Gkountela, J. Landin, M. C. Scheidmann, C. Donato, R. Scherrer, J. Singer, C. Beisel, C. Kurzeder, V. Heinzelmann-Schwarz, C. Rochlitz, W. P. Weber, N. Beerenwinkel, N. Aceto, Neutrophils escort circulating tumour cells to enable cell cycle progression. *Nature*. **566**, 553–557 (2019).

22. A. Giladi, M. Cohen, C. Medaglia, Y. Baran, B. Li, M. Zada, P. Bost, R. Blecher-Gonen, T. M.

Salame, J. U. Mayer, E. David, F. Ronchese, A. Tanay, I. Amit, Dissecting cellular crosstalk by sequencing physically interacting cells. *Nat. Biotechnol.* **38** (2020), doi:10.1038/s41587-020-0442-2.

23. N. Andrews, J. T. Serviss, N. Geyer, A. B. Andersson, E. Dzwonkowska, I. Šutevski, R. Heijboer, N. Baryawno, M. Gerling, M. Enge, An unsupervised method for physical cell interaction profiling of complex tissues. *Nat. Methods 2021*, 1–9 (2021).

24. G. Peng, S. Suo, G. Cui, F. Yu, R. Wang, J. Chen, S. Chen, Z. Liu, G. Chen, Y. Qian, P. P. L. Tam, J. D. J. Han, N. Jing, Molecular architecture of lineage allocation and tissue organization in early mouse embryo. *Nature*. **572**, 528–532 (2019).

25. D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, I. Amit, Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science (80-. ).* **343**, 776–779 (2014).

26. H. Keren-Shaul, E. Kenigsberg, D. A. Jaitin, E. David, F. Paul, A. Tanay, I. Amit, MARS-seq2.0: an experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing. *Nat. Protoc.* **14**, 1841–1862 (2019).

27. C. S. McGinnis, L. M. Murrow, Z. J. Gartner, DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* **8**, 329–337 (2019).

28. G. Peng, M. Westerfield, Lhx5 promotes forebrain development and activates transcription of secreted Wnt antagonists. *Development*. **133**, 3191–3200 (2006).

29. S. J. B. Butt, V. H. Sousa, M. V. Fuccillo, J. Hjerling-Leffler, G. Miyoshi, S. Kimura, G. Fishell, The Requirement of Nkx2-1 in the Temporal Specification of Cortical Interneuron Subtypes. *Neuron*. **59**, 722–32 (2008).

30. J. A. Belo, T. Bouwmeester, L. Leyns, N. Kertesz, M. Gallo, M. Follettie, E. M. De Robertis, Cerberus-like is a secreted factor with neuralizing activity expressed in the anterior primitive endoderm of the mouse gastrula. *Mech. Dev.* **68**, 45–57 (1997).

31. E. Hermesz, S. Mackem, K. A. Mahon, Rpx: a novel anterior-restricted homeobox gene progressively activated in the prechordal plate, anterior neural plate and Rathke's pouch of the mouse embryo. *Development*. **122**, 41–52 (1996).

32. K. Daneshvar, J. V. Pondick, B. M. Kim, C. Zhou, S. R. York, J. A. Macklin, A. Abualteen, B. Tan, A. A. Sigova, C. Marcho, K. D. Tremblay, J. Mager, M. Y. Choi, A. C. Mullen, DIGIT Is a Conserved Long Noncoding RNA that Regulates GSC Expression to Control Definitive Endoderm Differentiation of Embryonic Stem Cells. *Cell Rep.* **17**, 353–365 (2016).

33. M. Hansson, D. R. Olesen, J. M. L. Peterslund, N. Engberg, M. Kahn, M. Winzi, T. Klein, P. Maddox-Hyttel, P. Serup, A late requirement for Wnt and FGF signaling during activin-induced formation of foregut endoderm from mouse embryonic stem cells. *Dev. Biol.* **330**, 286–304 (2009).

34. M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, A. Ma'ayan, Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).

35. V. Barone, M. Lang, S. F. G. Krens, S. J. Pradhan, S. Shamipour, K. Sako, M. Sikora, C. C. Guet, C. P. Heisenberg, An Effective Feedback Loop between Cell-Cell Contact Duration and Morphogen Signaling Determines Cell Fate. *Dev. Cell*. **43**, 198–211 (2017).

36. E. Hannezo, C. P. Heisenberg, Mechanochemical Feedback Loops in Development and Disease. *Cell*. **178**, 12–25 (2019).

37. N. Perrimon, C. Pitsouli, B. Z. Shilo, Signaling mechanisms controlling cell fate and embryonic patterning. *Cold Spring Harb. Perspect. Biol.* **4** (2012), doi:10.1101/cshperspect.a005975.

38. C. P. Heisenberg, Y. Bellaïche, Forces in tissue morphogenesis and patterning. *Cell*. **153** (2013), p. 948.

39. N. I. Petridou, S. Grigolon, G. Salbreux, E. Hannezo, C. P. Heisenberg, Fluidization-mediated

tissue spreading by mitotic cell rounding and non-canonical Wnt signalling. *Nat. Cell Biol.* **21**, 169–178 (2019).

40. A. Miquelajáuregui, T. Sandoval-Schaefer, M. Martínez-Armenta, L. Pérez-Martínez, A. Cárabez, Y. Zhao, M. Heide, G. Alvarez-Bolado, A. Varela-Echavarría, LIM homeobox protein 5 (Lhx5) is essential for mamillary body development. *Front. Neuroanat.* **9**, 136 (2015).

41. O. Cleaver, P. A. Krieg, Notochord patterning of the endoderm. *Dev. Biol.* **234** (2001), pp. 1–12.

42. C. G, P. MT, S. D, K. J, N. M, H. P, X. MH, G. A, B. S, L. XH, C. K, B. M, F. N, G. HP, ANGPTL3 stimulates endothelial cell adhesion and migration via integrin alpha vbeta 3 and induces blood vessel formation in vivo. *J. Biol. Chem.* **277**, 17281–17290 (2002).

43. O. R. Rodrigues, S. Monard, A rapid method to verify single-cell deposition setup for cell sorters. *Cytom. Part A*. **89**, 594–600 (2016).

44. D. Kim, B. Langmead, S. L. Salzberg, HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods*. **12**, 357–360 (2015).

45. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive Integration of Single-Cell Data. *Cell*. **177**, 1888–1902 (2019).

46. L. Van Der Maaten, G. Hinton, Visualizing Data using t-SNE. *J. Mach. Learn. Res. 1*. **620**, 267–84 (2008).