Detecting Cell Contact-dependent Gene Expression from Spatial Transcriptomics Data

Hyobin Kim[1]*, Cecilia Lövkvist[1]*, Patrick Martin[1], Junil Kim[2] and Kyoung Jae Won[1]**

[1] Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark

[2] School of Systems Biomedical Science, Soongsil University, 369 Sangdo-Ro, Dongjak-Gu, Seoul, Republic of Korea

* Equally contributed to this work

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

Abstract

Understanding the cellular microenvironment promises to deliver insights into disease and development, yet we still lack the ability to systematically leverage high resolution spatial transcriptomics to understand cell-cell interactions. With the aim of identifying cell-contact dependent genes, we propose a novel approach that explores the transcriptome of spatial beads composed of multiple cell types (heterotypic beads). Our approach identifies cell-contact dependent expression beyond what ligand-receptor co-expression can identify.

Keywords: cell contact-dependent expression, cell-cell interactions, heterotypic/homotypic spots, spatial transcriptomics data

Main

Cell-cell interactions play a crucial role in conducting cellular functions in a given tissue architecture. Cells communicate with their microenvironment in various ways including release of soluble molecules and direct cell contacts[1], actively changing their transcriptomes in response to external signals[2, 3].

Recent development on spatial transcriptomics (ST) has opened potential ways to explore the role of the microenvironment as the transcriptome of a cell is measured together with that of the neighborhood. In contrast to image-based ST approaches such as MERFISH and seqFISH, the next-generation-sequencing (NGS) based ST technologies[4-6] do not require pre-selection of genes and can profile the transcriptome in an unbiased way. The recent technological advances in ST[4-6] have improved the spatial resolution to a near cellular level. Nevertheless, there is no systematic way to study gene expression associated with microenvironment at cellular resolution. Current approaches only focus on frequently interacting cell type groups based on ligand-receptor pairs [7, 8].

Analysis for NGS-based ST approaches is not trivial as a bead or a spot used in these techniques does not perfectly capture a cell, it can thereby contain transcriptomes of multiple cells. To study cell-contact dependent gene expression, we capitalized on spots with multiple cell types as evidence for cell-cell contacts. For instance, Slide-seq, a ST approach with 10 μm resolution, has *homotypic* (the same cell type in a bead) and *heterotypic* (two or more cell types in a bead) spots and the majority of the heterotypic spots are composed of two cell types (97% of spots are composed of one or two cell types)[4]. We used the heterotypic spots in Slide-seq.

We hypothesized that genes expressed at higher levels in heterotypic spots than homotypic spots for each cell type are evidence for cell-contact dependent gene expression (Fig. 1a and Methods). To find heterotypic spots, we applied the RCTD[2] deconvolution tool against Slide-seq V2 data, using single cell RNA-sequencing (scRNA-seq) data for mouse embryo and hippocampus[9, 10] as references. Then, we found differentially expressed genes (DEGs) between the heterotypic spots and the homotypic spots (Student's t-test, Welch's t-test, Mann-Whitney U test: log2(fold-change) > 0.8 and p-value < 0.01). We further confirmed the statistical significance (p-value < 0.001) of the DEGs generating a null model of artificial heterotypic spots. We created them by randomly selecting homotypic spots and mixing their transcriptomes according to cell type proportions in true heterotypic spots (Methods and Supplementary Fig. 1).

The RCTD deconvolution approach identified 8,094 homotypic and 34,268 heterotypic spots in the mouse embryo and 12,013 homotypic and 29,331 heterotypic spots in the mouse hippocampus. Using our workflow, we found 19 DEGs from 9 heterotypic pairs in the embryo and 90 DEGs from 21 heterotypic pairs in the hippocampus (Supplementary Fig. 2 and Supplementary Table 1).

For instance, we identified 7 genes including *Crim1* and *Cd24a* more highly expressed in the heterotypic spots of Endothelial and Lens cells compared with their respective homotypic spots in the embryo Slide-seq V2 data (Fig. 1b, top-left). The heterotypic spots contain the expression of marker genes for both Endothelial (*Pecam1* and *Egfl7*) and Lens cells (*Cryba1* and *Cryaa*). To confirm this, we visualized the cell-contact specific genes we obtained for both heterotypic and homotypic spots. In the spatial mapping, *Cd24a* is highly expressed in the heterotypic spots for Endothelial+Lens cells (red circle), but not for homotypic spots for Endothelial (blue) or Lens (black) cells (Fig. 1b, right). Intriguingly, the *Cd24a* (CD24) protein is known as a ligand for P-selectin, an adhesion molecule present in activated platelets and endothelial cells[11]. Investigating the scRNA-seq data taken from embryo[9], we found that *Cd24a* is mostly expressed in Endothelial cells (Fig. 1b, bottom-left, Supplementary Fig. 3 and Supplementary Table 1).

In the hippocampus Slide-seq V2 data, we identified 13 genes highly expressed in the heterotypic spots of Entorhinal cortex cells and Ependymal cells (Fig. 1c, top-left). From the scRNA-seq data taken from hippocampus[10], we examined potential cell-types expressing these genes (Fig. 1c, bottom-left). *Gabbr1*, for example, is expressed in Entorhinal cortex cells when they are proximal to Ependymal cells. Its spatial visualization shows that *Gabbr1* is more expressed in the heterotypic spots than the homotypic ones (Fig. 1c, right). *Gabbr1* is known as a receptor for gamma-aminobutyric acid (GABA), a neurotransmitter in the mammalian central nervous system[12].

Our findings suggest that cells express specific sets of genes depending on neighboring cell types. We found that Melanocytes (Ml) express *Col9a1* when next to Lens (L) and *Sfrp2* when next to Definite Erythroid Lineage (DEL) in the embryo (Supplementary Table 1). To investigate neighboring cell-type dependent gene expression, we colored the boundary of spots of Ml based on its neighboring cell types: red when proximal to L, green when to DEL, and orange when to another Melanocyte. To easily distinguish gene expression change depending on neighboring cell types, we defined the neighboring cell type specific genes by RGB color channels (see Methods): *Col9a1* (red) and *Sfrp2* (green) and represented the expression of these two genes using the combination of each color channel. Between the two genes, *Col9a1* (red inside the boundary) is dominantly expressed when Ml are proximal to L (Ml+L, red boundary) while *Sfrp2* (green inside the boundary) is dominantly expressed when Ml are proximal to DEL (Ml+DEL, green boundary) (Fig. 2a). We confirmed that not L and DEL but Ml mostly express these two genes in the scRNA-seq data from embryo[9] (Fig. 2b). These results indicate that Ml vary the expression levels of the two genes depending on the neighboring cell type of Ml. In the hippocampus, we also observed similar neighboring cell type dependent expression changes: *Scube1* (red) and *Tac2* (green) in Interneurons (In) when contacting Cajal-Retzius (Ca) and Choroid (Ch), respectively (Fig. 2c, 2d).

Ligand-receptor co-expression has been used to study cell-cell interactions. To see if the use of ligand-receptor pairs and the downstream genes mediated by them can recover our DEG set, we ran NicheNet[8] on the homotypic spots in Slide-seq V2 (see Methods, Supplementary Table 2). We found only one and 5 genes commonly detected between the two approaches in the embryo and hippocampus, respectively (Fig. 2e, Supplementary Table 3). To see if they are valid interactions, we calculated minimum distances between the interacting two cell types identified by NicheNet using the spatial coordinates of Slide-seq V2 (Supplementary Fig. 4). The signaling range was estimated from 50 to 2500 μm on average. This suggests that interactions identified by NicheNet may include many false predictions. It is also of note that NicheNet only shows interacting cell types and does not examine interactions between individual cells.

To demonstrate the usefulness of using the heterotypic spots in Slide-seq V2, we additionally ran NicheNet on the heterotypic spots. We set the heterotypic spots as a receiver as well as a sender (autocrine mode in NicheNet). We found that 2 genes (of 19 DEGs from 9 heterotypic pairs) in the embryo and 11 genes (of 90 DEGs from 21 heterotypic pairs) in the hippocampus common to our DEGs (Supplementary Table 4). For instance, *Sfrp2*, an inhibitor of Wnt signaling[13], was found as a ligand in the heterotypic spots of DEL+Melanocytes in the embryo. These results suggest that the use of heterotypic spots in ST dataset can be useful to study cell-cell communication.

To understand transcriptomic changes associated with cell direct contact, we investigated the highly expressed genes in the heterotypic spots in Slide-Seq V2. Gene expression changes due to cell-contact have previously been observed[14]. However, they have just focused on interactions between predefined cell types. Hence, there has been no systematic way to obtain them. The use of heterotypic spots enabled identification of cell-contact specific genes in an unbiased way beyond previous approaches that use ligand-receptor pairs. It provides opportunities to study cell-cell interactions between two adjacent cells from a new perspective and further identify new genes potentially involved in intercellular communication.

Methods

Slide-seq V2 data pre-processing. We used Slide-seq V2 data[5] in mouse embryo and hippocampus that are at near-cellular resolution (*i.e.*, 10 μm). The embryo and hippocampus datasets were obtained from Puck_190926_03 and Puck_200115_08, respectively. Each dataset has a count matrix (genes × barcoded beads) and information on X-Y coordinates of the beads. Quality control, normalization, and identification of highly variable features were performed with Seurat 3.2.2. For the embryo, samples that have unique feature count less than 200 were filtered out. For the hippocampus, all samples were used without the filtering because samples with unique feature counts less than 200 take up a considerable percentage of the total samples (i.e., about 40%). The values of the count matrix per dataset were log-normalized and then the top 2000 variable genes as well as cell type markers (Supplementary Table 5) were selected.

Cell type inference of Slide-seq spots. We used RCTD[2] (robust cell type decomposition) to identify the cell types of the spots in Slide-seq. RCTD is a deconvolution tool to decompose the cell type mixture of individual spots in spatial transcriptomics data, particularly for Slide-seq V2. It leverages scRNA-seq data where cell types are annotated. We performed the simulation under doublet mode[2] that constraints each spot to contain up to two cell types, which is recommended for data with fine resolution such as Slide-seq. If a spot consists of one cell type, it is defined as *homotypic*. If a spot consists of two different cell types, it is defined as *heterotypic*. To run RCTD on the embryo and hippocampus data, we trained RCTD with publicly available scRNA-seq datasets with annotated cell types. For embryo, we used a scRNA-seq dataset[9] at E12.5 equivalent to the developmental stage of Slide-seq embryo. It consists of 26,183 genes and 270,197 cells assigned into 37 cell types (Supplementary Fig. 5a). For hippocampus, we obtained a scRNA-seq dataset[10] from DropViz. It is composed of 27,953 genes and 113,507 cells assigned into 17 cell types (Supplementary Fig. 5b). Assuming that the cell type annotations in the scRNA-seq datasets are sufficiently accurate, we trained RCTD to predict the cell types of spots in Slide-seq. Afterwards, the results on the inference of cell types were validated using cell type markers (Supplementary Table 5) accompanied with the scRNA-seq datasets[9, 10]. The simulation was conducted in R.

Cell contact-dependent gene expression analysis. We studied cell contact-dependent gene expression by comparing heterotypic spots (two different cell types in a spot) with homotypic spots (one cell type in a spot). We identified genes that are more expressed when two different cell types are interacting with each other. We carried out rigorous statistical analysis between the two samples (*i.e.*, heterotypic spots and homotypic spots). We determined whether to use parametric or non-parametric two-sided tests depending on sample size. When both samples were larger than sample size 30, we chose parametric tests under the normality assumption. Specifically, we conducted the Student's t-test for equal variances and the Welch's t-test for unequal variances, where a two-sample F-test was used to test if the variances are equal. Meanwhile, when the sample size of at least one sample was smaller than 30, we performed the Mann-Whitney U test as a non-parametric test. The log2(fold-change) > 0.8 and p-value < 0.01 were used as criteria for differential expression. For example, when we had heterotypic spots of which cell type is A+B, we compared them with the homotypic spots of A and the homotypic ones of B, respectively. Genes that satisfied the criteria in both comparisons were finally identified as differentially expressed genes (DEGs). Matlab R2021b was used to perform cell contact-dependent gene expression analysis.

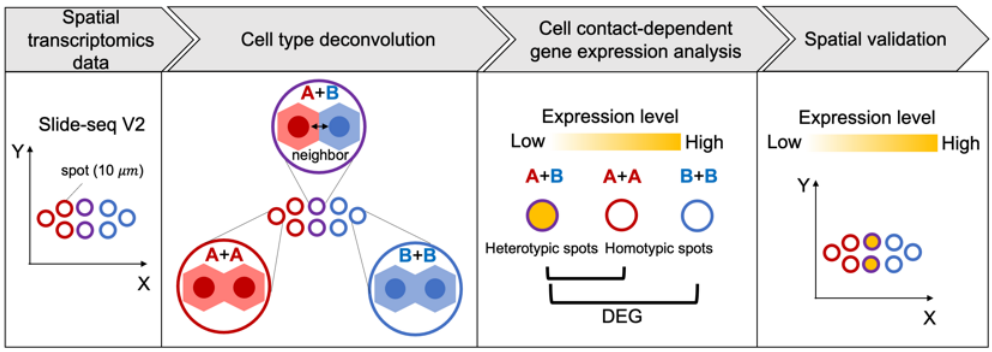
Verification of cell-cell interactions in heterotypic spots. We developed a null model to verify that individual heterotypic spots in Slide-seq represent two different cell types interacting with each other (Supplementary Fig. 1). Our null model refers to artificial heterotypic spots. We generated them by combining two homotypic spots based on the cell-type proportions of heterotypic spots obtained from RCTD. For example, a heterotypic spot whose cell type is A+B comprises α% of A and β% of B. One spot among the homotypic spots of A and another among the homotypic spots of B are randomly chosen, and then the respective transcriptomes are taken by α from the homotypic spot of A and by β from the homotypic one of B. By combining them, an artificial heterotypic spot with α% of A and β% of B is

produced. It represents two different cell types just combined without intercellular interaction. With repeated random sampling, we created 100 artificial heterotypic spots for each heterotypic spot. We compared the artificial heterotypic spots with the heterotypic spots of Slide-seq to find genes that are highly expressed in the latter. The same statistical tests as cell-contact dependent gene expression analysis were applied. The log2(fold-change) > 0.8 and p-value < 0.001 were used as criteria for differential expression. Our null model was implemented in Matlab R2021b.

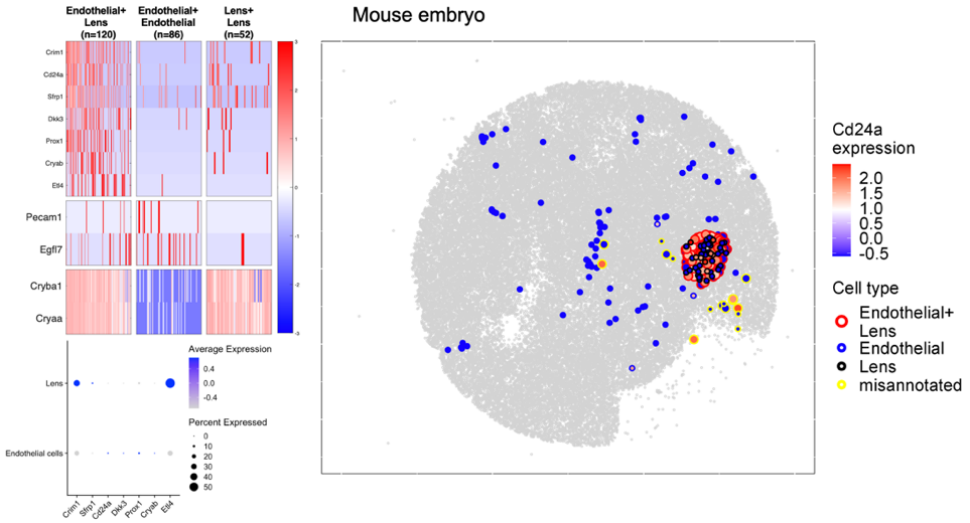
Spatial visualization with RGB color channels. We used RGB coordinates (r, g, b) composed of integer values from 0 to 255 to study how gene expression varies depending on neighboring cell types. We first selected one cell type and then collected heterotypic pairs where the selected cell type is included. For example, the cell type of interest is A, and heterotypic pairs A+B and A+C are collected. If *gene* γ and *gene* δ are DEGs found from the heterotypic spots of A+B and A+C, respectively, red and green channels are assigned to the expression values of the two DEGs: *gene* γ is red (R) and *gene* δ is green (G). The value of the blue channel is fixed as zero. The expression of the DEGs is simultaneously visualized. How much red and green are mixed produces diverse colors. For example, if the RGB color is black, it means that the two genes are not expressed at all. If the RGB color is red (or green), it indicates that the expression of *gene* γ (or *gene* δ) is dominant, but the other is not. The spatial visualization with RGB color channels was performed in R.

Identification of ligands, receptors, and downstream targets. We used NicheNet[8] to compare our DEGs with already known ligands, receptors, and their target genes. NicheNet predicts which ligands in sender cells affect which receptors and target genes in receiver cells based on prior knowledge on ligand-to-target signaling paths in humans and mice. We ran NicheNet on Slide-seq data in two ways: (i) To see if NicheNet can detect our DEGs, we investigated homotypic spots that correspond to the heterotypic pairs (9 heterotypic pairs in the embryo and 21 heterotypic pairs in the hippocampus) where our DEGs were identified. We set the respective homotypic spots as a sender and a receiver by turns. (ii) To validate if the heterotypic spots in Slide-seq are useful to study cell-cell interactions, we examined the heterotypic spots where our DEGs were identified. We set the heterotypic spots as a receiver as well as a sender (autocrine mode in NicheNet). We found highly expressed ligand-receptor-target genes, where highly expressed genes mean expressed in at least 10% of cells in one cluster.

a.



b.



c.

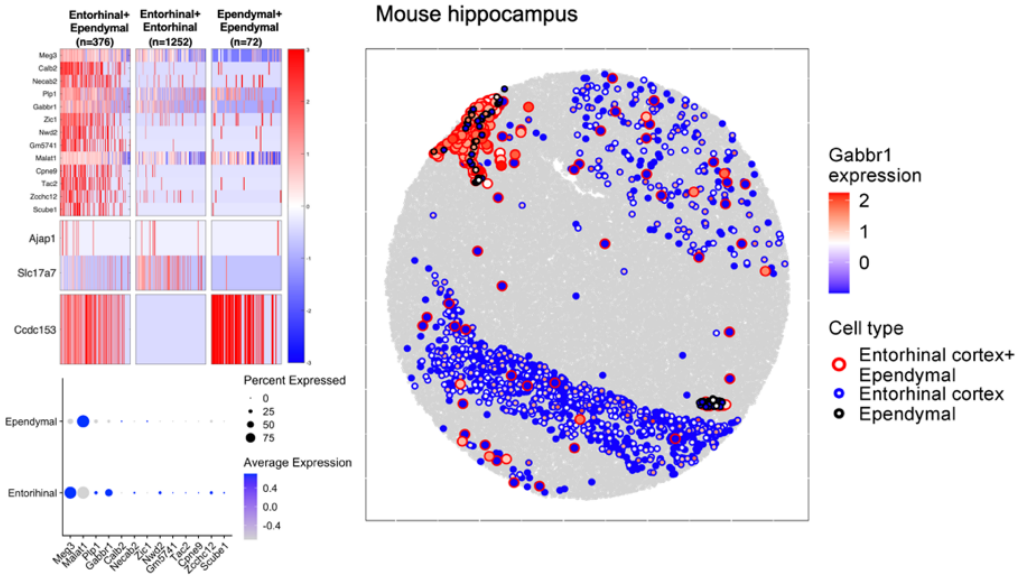


Figure 1. Cell contact-dependent gene expression analysis. a, Workflow for cell contact-dependent gene expression analysis. **b**, Cell-contact specific genes for Endothelial and Lens cells in mouse embryo. **c**, Cell-contact specific genes for Entorhinal cortex cells and Ependymal cells in mouse hippocampus.

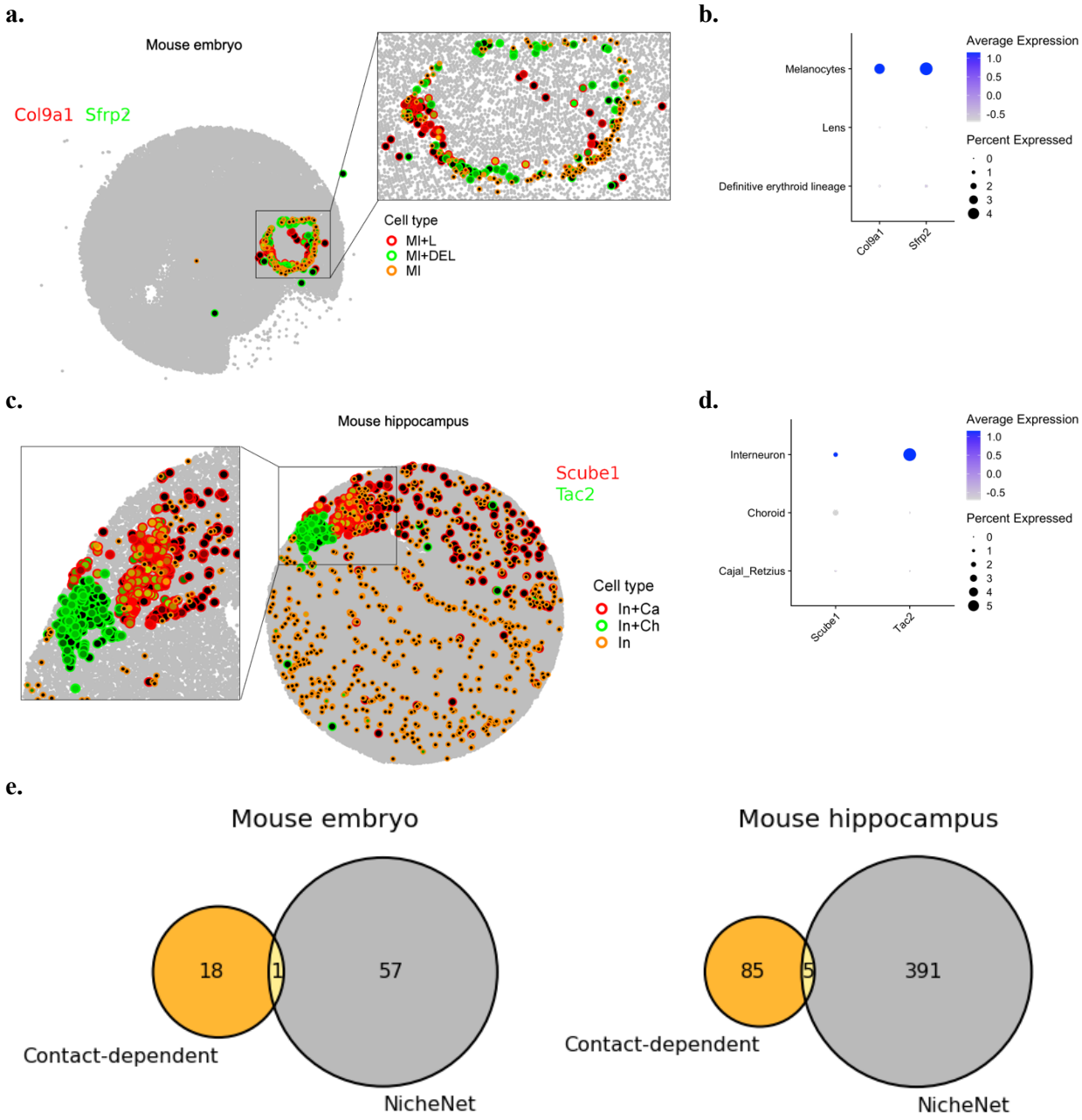


Figure 2. Neighboring cell-type dependent gene expression. a, Spatial mapping with RGB channels of contact specific genes in mouse embryo. Melanocytes (Ml) dominantly express *Col9a1* (red) when proximal to Lens (L), and *Sfrp2* (green) when proximal to Definite Erythroid Lineage (DEL). **b**, Expression of *Col9a1* and *Sfrp2* in Ml, L, and DEL in the scRNA-seq data from embryo. **c**, Spatial mapping with RGB channels of contact specific genes in mouse hippocampus. Interneurons (In) dominantly express *Scube1*(red) when proximal to Cajal-Retzius (Ca), and *Tac2* (green) when proximal to Choroid (Ch). **d**, Expression of *Scube1* and *Tac2* in In, Ch, and Ca in the scRNA-seq data from hippocampus. **e,** Venn diagrams of genes identified between cell contact-dependent gene analysis and NicheNet.

Data availability

Slide-seq V2 data in mouse embryo and hippocampus are available at https://singlecell.broadinstitute.org/single_cell/study/SCP815/sensitive-spatial-genome-wide-expression-profiling-at-cellular-resolution#study-summary.

Code availability

Codes for the cell contact-dependent gene expression analysis and spatial visualization are available at https://github.com/hkim240/CellContact. RCTD is available at https://github.com/dmcable/RCTD. NicheNet is available at https://github.com/saeyslab/nichenetr.

References

1. Yang, B.A. et al. Engineered Tools to Study Intercellular Communication. *Advanced Science* **8**, 2002825 (2021).
2. Cable, D.M. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 1-10 (2021).
3. Dries, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology* **22**, 1-31 (2021).
4. Rodriques, S.G. et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463-1467 (2019).
5. Stickels, R.R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature biotechnology* **39**, 313-319 (2021).
6. Liu, Y. et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell* **183**, 1665-1681. e1618 (2020).
7. Efremova, M. et al. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature protocols* **15**, 1484-1506 (2020).
8. Browaeys, R. et al. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nature methods* **17**, 159-162 (2020).
9. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).
10. Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015-1030. e1016 (2018).
11. Sammar, M. et al. Heat-stable antigen (CD24) as ligand for mouse P-selectin. *International immunology* **6**, 1027-1036 (1994).
12. Tillakaratne, N.J. et al. Gamma-aminobutyric acid (GABA) metabolism in mammalian neural and nonneural tissues. *Comparative Biochemistry and Physiology Part A: Physiology* **112**, 247-263 (1995).
13. Jin, L. et al. SFRP2 enhances the osteogenic differentiation of apical papilla stem cells by antagonizing the canonical WNT pathway. *Cellular & molecular biology letters* **22**, 1-14 (2017).
14. Küppers, M. et al. The transcriptional programme of contact-inhibition. *Journal of Cellular Biochemistry* **110**, 1234-1243 (2010).