Title
Screening cell–cell communication in spatial transcriptomics via collective optimal transport

Journal
Nature Methods, 20(2)

ISSN
1548-7091

Authors
Cang, Zixuan
Zhao, Yanxiang
Almet, Axel A
et al.

Publication Date
2023-02-01

DOI
10.1038/s41592-022-01728-4

Peer reviewed

nature methods

Article

https://doi.org/10.1038/s41592-022-01728-4

Screening cell–cell communication in spatial transcriptomics via collective optimal transport

Received: 8 October 2021

Accepted: 21 November 2022

Published online: 23 January 2023

Check for updates

Zixuan Cang[1], Yanxiang Zhao[2], Axel A. Almet[3,4], Adam Stabell[4,5], Raul Ramos[4,5], Maksim V. Plikus[4,5], Scott X. Atwood[4,5] & Qing Nie[3,4,5]

Spatial transcriptomic technologies and spatially annotated single-cell RNA sequencing datasets provide unprecedented opportunities to dissect cell–cell communication (CCC). However, incorporation of the spatial information and complex biochemical processes required in the reconstruction of CCC remains a major challenge. Here, we present COMMOT (COMMunication analysis by Optimal Transport) to infer CCC in spatial transcriptomics, which accounts for the competition between different ligand and receptor species as well as spatial distances between cells. A collective optimal transport method is developed to handle complex molecular interactions and spatial constraints. Furthermore, we introduce downstream analysis tools to infer spatial signaling directionality and genes regulated by signaling using machine learning models. We apply COMMOT to simulation data and eight spatial datasets acquired with five different technologies to show its effectiveness and robustness in identifying spatial CCC in data with varying spatial resolutions and gene coverages. Finally, COMMOT identifies new CCCs during skin morphogenesis in a case study of human epidermal development.

The complex structures and functions of multicellularity are achieved through the coordinated activities of various cells. Cells make decisions and accomplish their goals by interacting with an environment consisting of external stimuli and other cells. A major form of cell–cell interaction is cell–cell communication (CCC), mainly mediated by biochemical signaling through ligand–receptor binding that induces downstream responses that shape development, structure and function.

Traditionally, CCC studies were restricted to a few cell types and a small number of selected genes at the resolution of cell groups. Recently, the emergence of single-cell transcriptomics (that is, single-cell RNA sequencing, scRNA-seq) has enabled the examination of tissues at single-cell resolution at unprecedented genomic coverage[1]. Computational tools have been developed to estimate CCC activities from scRNA-seq data[2,3] using signaling databases[4–6]. Most of these methods rely on the expression levels of ligand and receptor pairs and explicitly defined functions. For example, the products of ligand and receptor levels[5,7] or non-linear Hill function-based models[6] are used. In addition, these methods emphasize different aspects of CCC. For example, CellPhoneDB[5], ICELLNET[7] and CellChat[6] account for the multi-subunit composition of protein complexes; SoptSC[8], NicheNet[9] and CytoTalk[10] utilize downstream intracellular gene–gene interactions; and scTensor[11] examines higher-order CCC represented as hypergraphs. These inference methods designed for scRNA-seq data have provided biological insights based on non-spatial transcriptomic data[2,12,13]. However, these non-spatial studies often contain

[1]Department of Mathematics and Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, USA. [2]Department of Mathematics, The George Washington University, Washington, DC, USA. [3]Department of Mathematics, University of California, Irvine, Irvine, CA, USA. [4]The NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine, Irvine, CA, USA. [5]Department of Developmental and Cell Biology, University of California, Irvine, Irvine, CA, USA. e-mail: qnie@uci.edu

significant false positives given that CCC takes place only within limited spatial distances that are not measured in scRNA-seq datasets. Improvements can be made by filtering the inferred CCC using spatial annotations[14].

Spatial transcriptomics[15–20] provides information on the distance between cells or spots containing multiple or fractions of cells. At various cellular resolutions these technologies measure the spatial expression of hundreds to tens of thousands of genes in 2-dimensional (2D) or 3-dimensional tissue (3D) samples[21]. Methods and software[22–24] developed for non-spatial data analysis have been applied to spatial data, with a small number of methods designed specifically for spatial data. Giotto builds a spatial proximity graph to identify interactions through membrane-bound ligand–receptor pairs[23]; CellPhoneDB v3 restricts interactions to cell clusters in the same microenvironment defined based on spatial information[25]; stLearn relates the co-expression of ligand and receptor genes to the spatial diversity of cell types[24]; SVCA[26] and MISTy[27] use probabilistic and machine learning models, respectively, to identify the spatially constrained intercellular gene–gene interactions; and NCEM fits a function to relate cell type and spatial context to gene expression[28]. However, current methods examine CCC locally and on cell pairs independently, and focus on information between cells or in the neighborhoods of individual cells. As a result, collective or global information in CCC, such as competition between cells, is neglected.

Optimal transport has recently been used for transcriptomic data analysis, including batch effect correction[29], developmental trajectory reconstruction[30] and spatial annotation of scRNA-seq data[31,32]. Naturally, one can form an optimal transport problem by viewing ligand and receptor expression as two distributions to be coupled with a cost based on spatial distance[31,33,34]. However, when using classical optimal transport, different molecule species with significantly different expression levels are normalized to ensure the same total mass, which renders the units of distributions unable to be compared. Furthermore, multiple ligand species can bind to multiple receptor species, resulting in competition. Of the 1,735 (secreted) ligand–receptor pairs in the Fantom5 database[35], 72% of ligands (372 of 516) and 60% of receptors (309 of 512) bind to multiple species. Such competition between multiple molecule species is ubiquitous and a critical biophysical process but it is ignored in existing methods. Although recent optimal transport variants such as unbalanced optimal transport and partial optimal transport can deal with unnormalized distributions and avoid certain coupling due to signaling spatial range and simultaneous consideration of multiple species[33,36–38], they introduce other issues. Specifically, unbalanced optimal transport[38] in its common form uses Kullback–Leibler divergence as a soft constraint on marginal distribution preservation. This approach may result in the total amount of coupled signaling molecule species significantly exceeding the total amount of either ligand or receptor initially available. By contrast, partial optimal transport[36] requires an additional parameter, the total coupled mass, which is usually difficult to estimate in the context of CCC inference.

To adapt optimal transport theory for the application of CCC inference, we present a method called collective optimal transport, which is capable of preserving the comparability between distributions, ensuring that the total signal does not exceed the individual species amounts (ligand or receptor), enforcing spatial range limits of signaling, and handling multiple competing species. The collective optimal transport method achieves this by optimizing the total transported mass and the ligand–receptor coupling simultaneously, unlike existing optimal transport methods. By introducing an entropy regularization to enforce the inequalities for marginal distributions, the collective optimal transport can be reformulated as a special case of the general unbalanced optimal transport framework[38]. An efficient algorithm is developed specifically for solving the collective optimal transport problem.

Based on collective optimal transport, we develop COMMunication analysis by Optimal Transport (COMMOT), a package that infers CCC by simultaneously considering numerous ligand–receptor pairs for either spatial transcriptomics data or spatially annotated scRNA-seq data equipped with spatial distances between cells estimated from paired spatial imaging data; summarizes and compares directions of spatial signaling; identifies downstream effects of CCC on gene expressions using ensemble of trees models; and provides visualization utilities for the various analyses.

We show that COMMOT accurately reconstructs CCC on simulated data generated by partial differential equation (PDE) models and outperforms three related optimal transport methods. We then apply COMMOT to analyze scRNA-seq data that have been spatially annotated using paired spatial datasets and five types of spatial transcriptomics data that differ with respect to spatial resolution or gene coverage. Finally, we examine a specific system of human epidermal development and elucidate connections between CCC and skin development.

Results

Overview of COMMOT

Ligands and receptors often interact with multiple species and within limited spatial ranges (Fig. 1a). Considering this, we present collective optimal transport (Fig. 1b) with three important features: first, the use of non-probability mass distributions to control the marginals of the transport plan to maintain comparability between species; second, enforcement of spatial distance constraints on CCC to avoid connecting cells that are spatially far apart; and last, the transport of multi-species distributions (ligands) to multi-species distributions (receptors) to account for multi-species interactions (Fig. 1c).

Given a spatial transcriptomics dataset of n_s cells or spots and n_l ligand species and n_r receptor species, the collective optimal transport determines an optimal multi-species coupling $\boldsymbol{P}^* \in \mathbb{R}_+^{n_l \times n_r \times n_s \times n_s}$ where $\boldsymbol{P}^*_{i,j,k,l}$ scores the signaling strength from sender cell k to receiver cell l through ligand i and receptor j. This is achieved by solving a minimization problem, $\min_{\boldsymbol{P}\in\Gamma} \sum_{(i,j)\in I} \alpha_{(i,j)} \langle \boldsymbol{P}_{i,j,\cdot,\cdot}, \boldsymbol{C}_{(i,j)} \rangle_F$ where

$$\Gamma = \left\{ \boldsymbol{P} \in \mathbb{R}_+^{n_l \times n_r \times n_s \times n_s} : \boldsymbol{P}_{i,j,\cdot,\cdot} = 0 \text{ for } (i,j) \notin I, \right.$$
$$\left. \textstyle\sum_{j,l} \boldsymbol{P}_{i,j,k,l} \le \boldsymbol{X}_{i,k}, \sum_{i,k} \boldsymbol{P}_{i,j,k,l} \le \boldsymbol{X}_{j,l} \right\},$$

I is the index set for ligand and receptor species that can bind together, and $\boldsymbol{X}_{i,k}$ is the expression level of gene i on spot k. The species-specific cost matrix $\boldsymbol{C}_{(i,j)}$ is a modified distance matrix for between-spot distance that replaces distances exceeding the spatial range of ligand i by infinity. The competitions between molecule species and cells are considered by assuming that a given receptor species or cell has limited capacity for interactions, such that a stronger inferred interaction with one ligand species or cell reduces the potential of interaction with other ligand species or cells (see the Methods and Supplementary Note for detailed formulations and algorithm derivations).

Direct validation of CCC inference methods for spatial data is difficult due to a lack of spatial co-localization measurements of ligand and receptor proteins. Here, we built PDE models to simulate CCC in space (Extended Data Fig. 1). Simulating various numbers of ligand and receptor species and diverse competition patterns, COMMOT accurately reconstructs the CCC connections from the resulting synthetic data (Extended Data Fig. 1d and Supplementary Figs. 1–4). COMMOT outperformed, and is significantly different from, two related optimal transport variants: unbalanced optimal transport and partial optimal transport (Supplementary Figs. 5–9). COMMOT's characteristics of enforcing spatial limits and not requiring probability distributions are further illustrated with other real spatial transcriptomics datasets (Supplementary Figs. 10 and 11).

https://doi.org/10.1038/s41592-022-01728-4

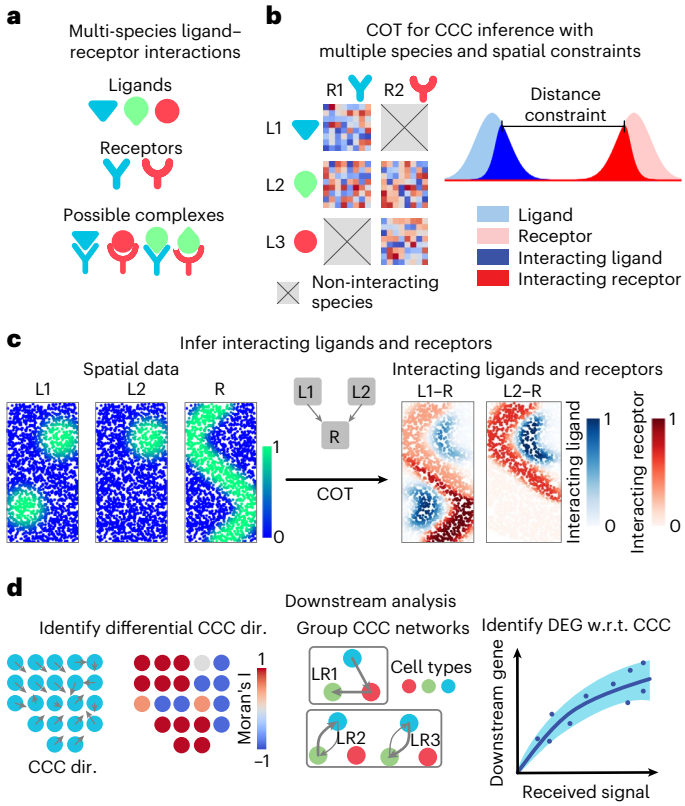


Fig. 1 | Overview of COMMOT. a, COMMOT infers CCC in space while considering the competition between different ligand and receptor species. **b**, Collective optimal transport (COT) infers CCC in space by introducing multi-species distributions and enforcing limited spatial ranges. **c**, An example of inferring CCC for spatial distributions of ligand–receptor complexes from spatial distributions of the ligands and receptor where two ligand species (L1, L2) compete for one receptor species (R). **d**, Three applications of downstream analysis based on the inferred CCC network between cells or spots. DEG, differentially expressed gene; dir., direction; w.r.t., with respect to.

For each ligand–receptor pair and each pair of cells or spots, the CCC inference quantifies the ligand contributed by one spot to the ligand–receptor complex in another spot. We then perform several downstream analyses: first, interpolation of the spatial signaling direction and identification of the differences between CCC regions; second, summarization and grouping of CCC at the spatial cluster level; and last, identification of the downstream genes affected by the CCC (Fig. 1d). The spatial signaling direction is obtained by interpolating the cell-by-cell CCC matrix to a vector field to identify the direction from which the signal is received or sent. For downstream analysis we first identify genes that are differentially expressed with the received signal, then quantify the CCC effect on these genes while considering the effect of other genes by incorporating a machine learning model that predicts a target gene level using both the received signal and other correlated genes. See Methods for the algorithms that perform the downstream tasks.

The roles of CCC in human epidermal development

We applied COMMOT to examine the development of epidermis in human skin. Our recent work profiled neonatal human epidermis using scRNA-seq and identified four stem cell clusters (basal I, II, III and IV) found in different regions of the innermost basal layer of the epidermis, a differentiating spinous cell cluster in the intermediate layer, and a granular cell cluster in the outermost living layers[39]. A refined in situ spatial transcriptomic map was constructed using SpaOTsc[31] by integrating scRNA-seq data with spatial data digitized from immunofluorescence staining images. The integrated dataset correctly identified previously known locations of the epidermal cell types and agreed with a known developmental path by epidermal cells from basal to suprabasal layers (Fig. 2a). This result was further validated by leave-one-out validation (Supplementary Fig. 12).

The spatial signaling between epidermal cells was inferred in the integrated dataset by considering ligand–receptor pairs annotated in the database CellChatDB. For example, our computational analysis predicted that molecular interactions between the ligands GAS6 and PROS1 with their receptor TYRO3 (GAS6-TYRO3 and PROS1-TYRO3) are significant in granular cells and moderately present in basal cells (Fig. 2b). This prediction was confirmed by both immunostaining for proteins (Fig. 2d) and using RNAscope to stain for RNA (Fig. 2e).

At the signaling pathway level we examined four specific pathways with known important roles in epidermal homeostasis, namely the WNT, TGF-β (transforming growth factor-β), NOTCH and JAK/STAT (Janus kinase/signal transducers and activators of transcription) pathways[39] (Fig. 2f and Supplementary Figs. 13–16). For all four pathways we observed mainly upward-directed signaling, with some downward signaling to the basal layers at the bottom of the ridges (Fig. 2f). WNT signaling is known to promote basal stem cell proliferation[40], whereas TGF-β suppresses it[41,42]. Thus, this observed directional signaling from the suprabasal layers may be regulating the communications to basal cells on proliferation.

Based on the inferred signaling activities, we further identified differentially expressed genes corresponding to each signaling pathway and modeled their expression level changes with increasing received signal without further considering spatial information (Fig. 2g). For the WNT pathway, increasing signal results in higher expression of the known basal cell markers *KRT15* and *KRT5*, as well as lower expression of the known terminally differentiated granular cell markers *LOR* and *FLG*, reinforcing the WNT pathway's known role in stem cell proliferation[40]. The analysis also predicted that higher WNT signaling would increase the expression of *BCAM*, *POSTN* and *STMN1*, the expression localization of which we confirmed by immunostaining on human epidermis (Fig. 2h). Interestingly, computational results predicted that *IGFBP6*, *PMAIP1* and *FGF7* would correlate positively with WNT signaling, but we observed their expression mainly in the spinous and granular layers, possibly due to predicted WNT signaling in both directions in basal-IV (Fig. 2h). TGF-β signaling had a similar profile to that of the WNT pathway, with NOTCH and JAK/STAT signaling having a more complex response (Fig. 2g). These results suggest how testable hypotheses can be derived from inferred signaling activities.

Signaling analysis in spatial transcriptomics data with high spatial resolution

We first studied CCC in spatial transcriptomics data with high spatial resolution using the CellChatDB[6]. We analyzed MERFISH (multiplexed error-robust fluorescence in situ hybridization) data of the mouse hypothalamic preoptic region with 161 genes and 73,655 cells across 12 slices along the anterior–posterior axis[43] (Fig. 3a–c). Of the signaling pathways available in the data, oxytocin (OXT) signaling, an important pathway that modulates social behaviors, was found to be most active. Self-modulation of excitatory neurons and modulation of inhibitory neurons by excitatory neurons through OXT signaling were identified across all of the slices (Fig. 3b, Extended Data Fig. 2 and Supplementary Fig. 17), a result consistent with the known major functions of OXT signaling[44]. Further analysis identified the local regions of high OXT signaling activity and the spatial direction of OXT signaling (Fig. 3c), which agreed with the results of protein staining of OXT and its receptor[45]. A gradual change of predicted signaling direction and high-activity regions was observed through adjacent slices (Fig. 3c and Extended Data Fig. 2).

https://doi.org/10.1038/s41592-022-01728-4

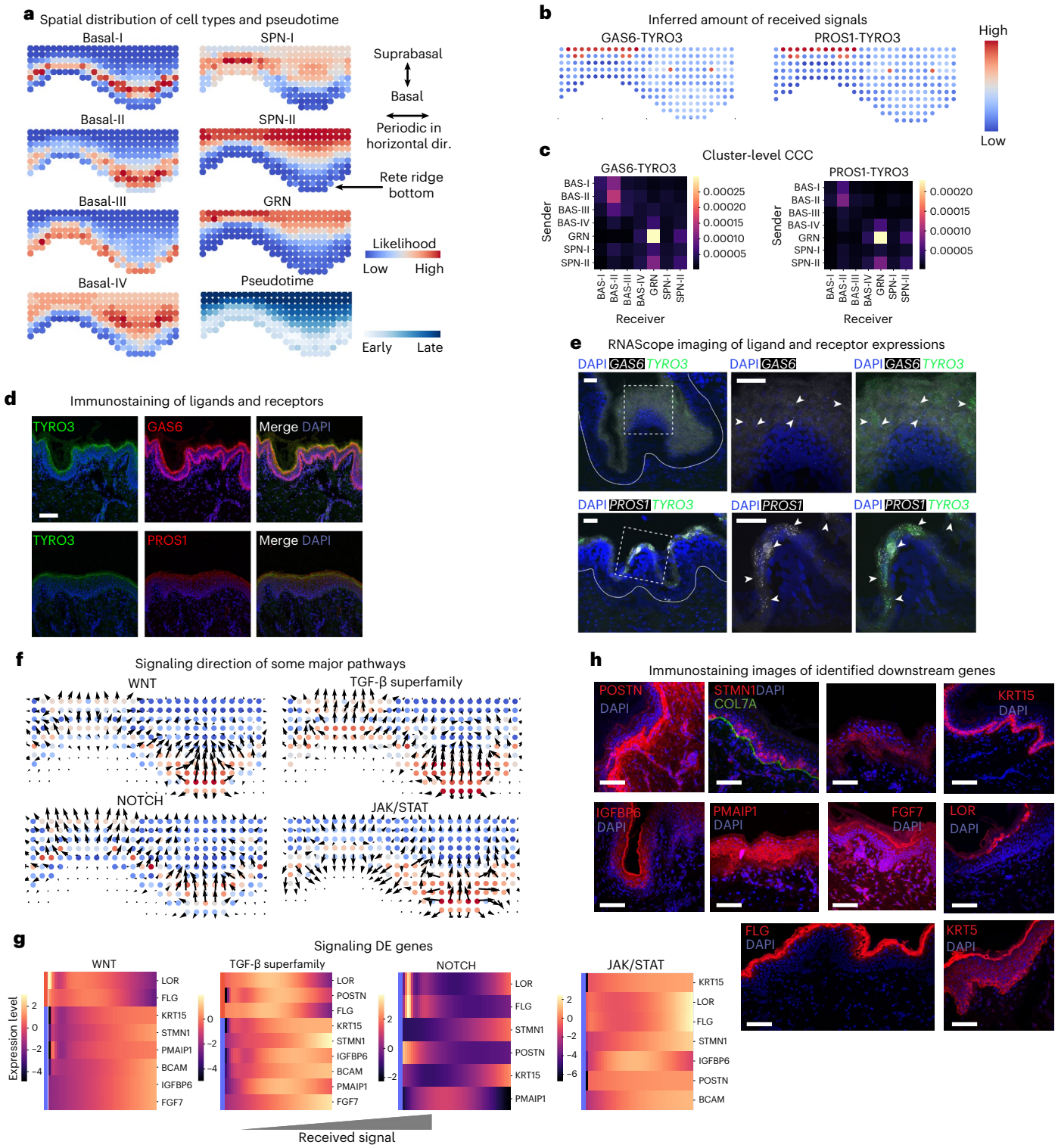


Fig. 2 | Role of CCC in human skin development. a, Predicted spatial origin of the skin subtypes of cells in intact tissue and the pseudotime projected to space. GRN, granular cell cluster; SPN, spinous cell cluster. **b**,**c**, The inferred amount of received signals of two example ligand–receptor pairs, GAS6-TYRO3 and PROS1-TYRO3 at the cell level (**b**) and cluster level (**c**). **d**, Immunostaining of proteins for GAS6, TYRO3 and PROS1. **e**, Fluorescent in situ hybridization against RNA molecules for predicted ligand–receptor interactions in human epidermis (solid white outline; regions of interest are marked by a white dashed square). The top row shows expression patterns of *GAS6* (white) and *TYRO3* (green); the bottom row shows expression patterns for *PROS1* (white) and *TYRO3* (green). In both cases, the middle and right panels show ligand–receptor signals, some of which colocalize to the stratum granulosum (white arrowheads). In merged images, the brightness of the *GAS6* channel was increased to improve clarity against the prominent *TYRO3* (green) signal. Experiments were repeated four times independently with consistent results. **f**, The signaling directions of four major signaling pathways. **g**, Heatmaps of selected signaling differentially expressed genes of the four signaling pathways, respectively. **h**, Immunofluorescence staining images of the identified signaling differentially expressed genes supporting the identified correlation between WNT signaling and the expression of these genes. Scale bars: **d**,**e**,**h**, 100 µm. The immunostaining experiments in **d** and **h** were repeated three times independently with consistent results.

https://doi.org/10.1038/s41592-022-01728-4

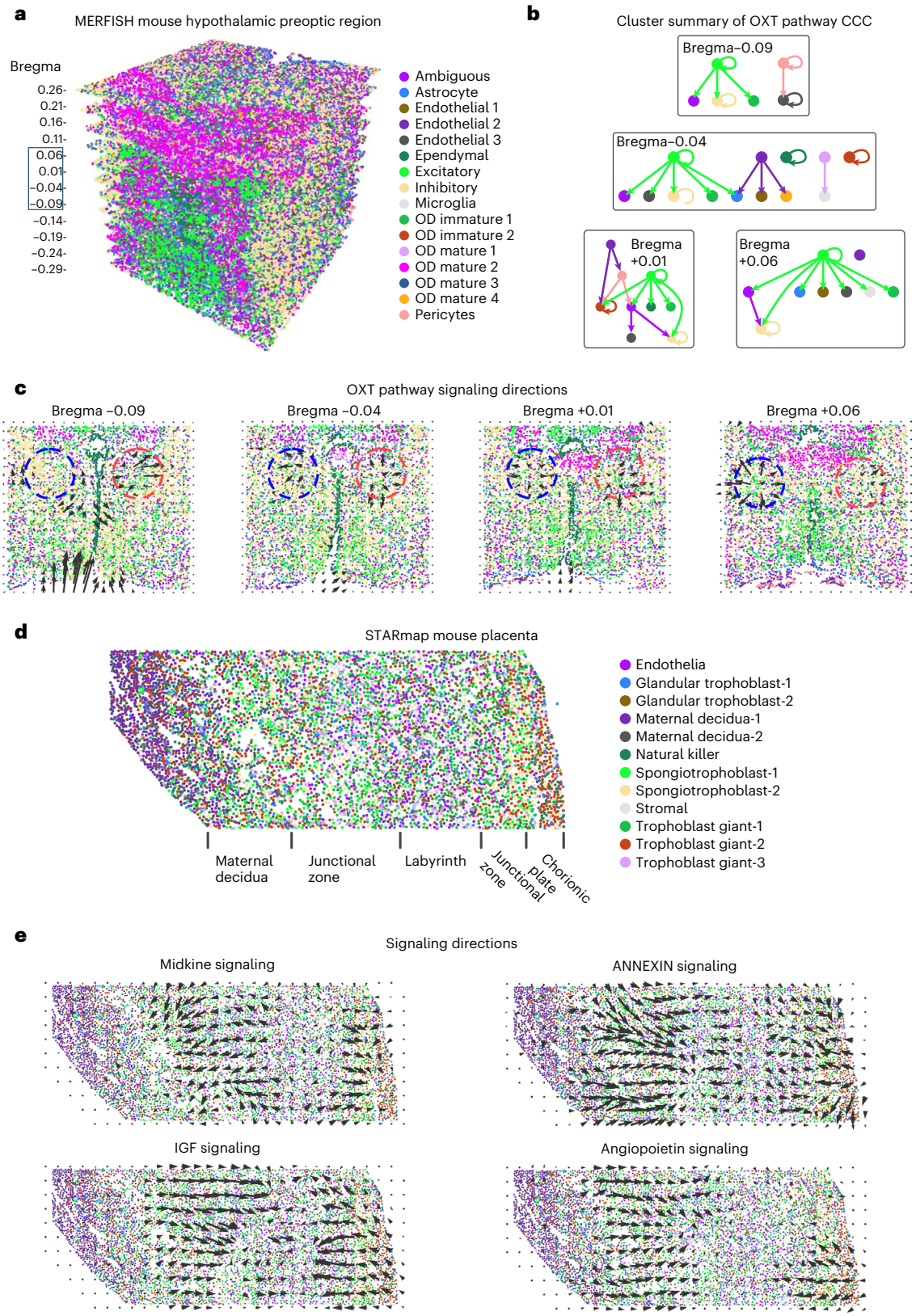


Fig. 3 | Inference of signaling direction in single-cell resolution spatial transcriptomics data. a, MERFISH data of the mouse hypothalamic preoptic region with multiple slices across the anterior–posterior axis[44]. **b**, Cluster-level summary of CCC through the OXT signaling pathway. **c**, Signaling directions of the OXT pathway. **d**, STARmap data of the mouse placenta[46]. **e**, Signaling directions of the midkine, IGF, annexin and angiopoietin pathways.

We then analyzed STARmap (spatially-resolved transcript amplicon readout mapping) data of mouse placenta with 903 genes and 7,203 cells[46] (Fig. 3d). Midkine and insulin-like growth factor (IGF) signaling were found to be active in the same regions but with opposing directions (Fig. 3e), suggesting a potential feedback loop[47]. In addition, it was found that IGF signaling is active in the labyrinth region and in endothelial

https://doi.org/10.1038/s41592-022-01728-4

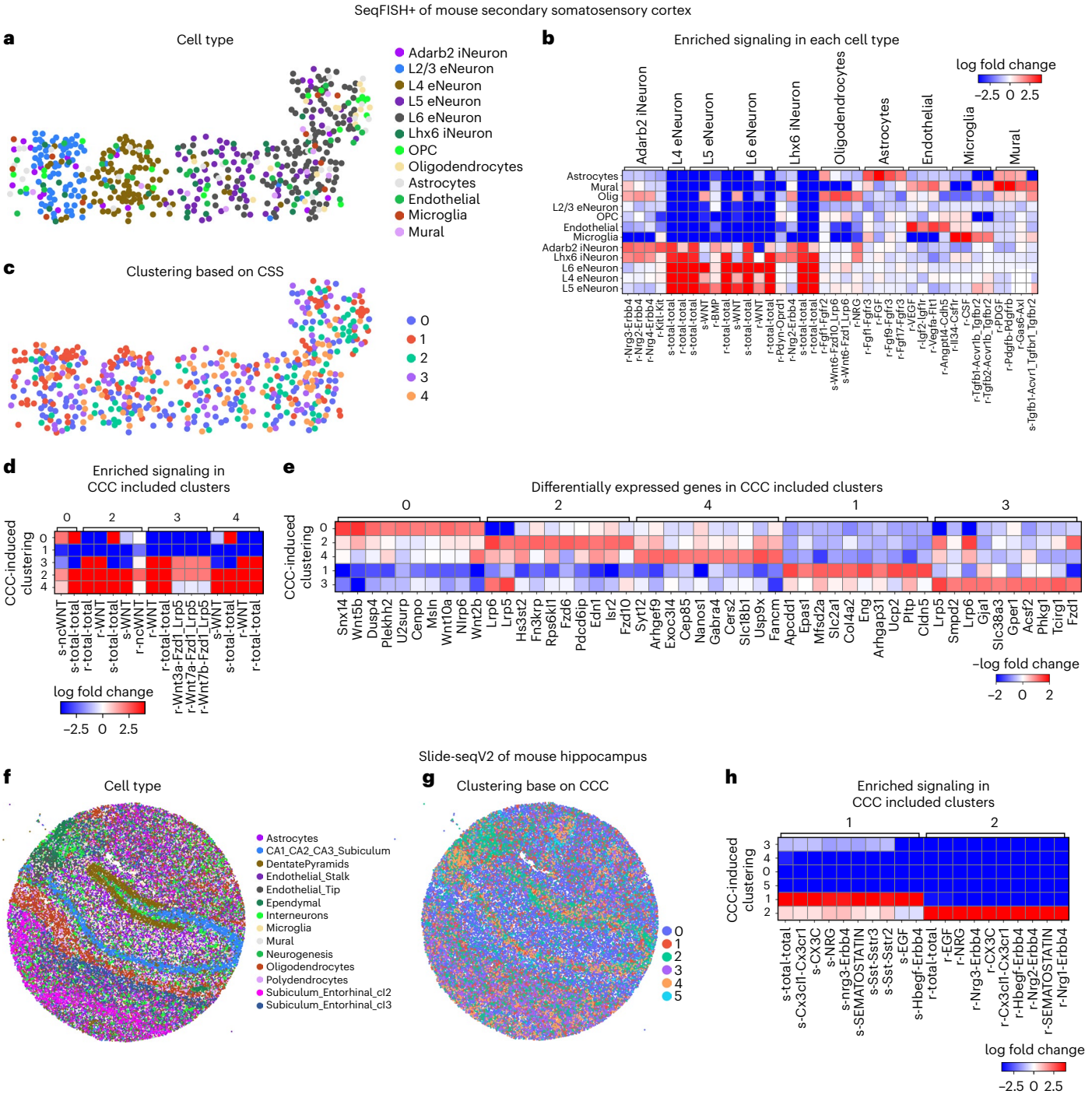


Fig. 4 | Downstream analysis of inferred CCC in single-cell resolution spatial transcriptomics data. a–**e**, CCC analysis of seqFISH+ data of mouse secondary somatosensory cortex. **a**, Clustering of cell type based on gene expression. OPC, oligodendrocyte precursor cells. **b**, Enriched signaling in each cell type. **c**, Clustering based on inferred CCC. **d**, Enriched signaling in CCC-induced clusters. **e**, Differentially expressed genes in the CCC-induced clusters. **f**–**h**, CCC analysis of Slide-seq (v2) data of mouse hippocampus. **f**, Clustering of cell type based on gene expression. **g**, Clustering based on inferred CCC. **h**, Enriched signaling in CCC-induced clusters.

cells, both of which were consistent with our predictions[48]. Midkine signaling was inferred to be active in trophoblast cells, consistent with previous findings on the role of SDC1 and SDC4 in trophoblast cells[49,50] (Supplementary Fig. 18). We also found that the annexin and the angiopoietin signaling pathways were active in similar regions with similar directions, suggesting that they may function cooperatively (Fig. 3e).

To demonstrate downstream analyses of CCC, we first studied seqFISH+ (sequential fluorescence in situ hybridization) data of mouse secondary somatosensory cortex with 10,000 genes measured in 523 individual cells[18] (Fig. 4a–e). Using the inferred CCC, each cell was assigned a CCC profile quantifying the amount of signal sent or received through each ligand–receptor pair to assemble a ($n_s \times 2n_{lr}$) CCC profile matrix for the n_s cells and n_{lr} ligand–receptor pairs. Differential expression analysis of the cell types and CCC profile found neuron cells to be most active through various ligand–receptor pairs, and distinct CCC activities for relatively rare cell types (Fig. 4b). Predicted

significant WNT signaling in neurons (Supplementary Fig. 19) correlated well with known critical roles of WNT signaling in neuronal migration and activity in the somatosensory cortex[51].

After clustering with respect to CCC activities, cells in the same group are expected to have similar signaling activities (Fig. 4c). Clusters 2 and 4 showed hyperactive signaling while clusters 0 and 3 were significant signal senders and receivers, respectively (Fig. 4d). We next identified differentially expressed genes that matched the signaling patterns of each CCC-induced cluster (Fig. 4e). This analysis identified both known signaling components in the relevant pathways and regulators of each pathway. For example, the positive differentially expressed genes associated with cluster 0 (WNT signal senders) included the known WNT ligands *Wnt5b*, *Wnt10a* and *Wnt2b*, while the differentially expressed genes in cluster 3 (WNT signal receivers) included known target genes of the WNT signaling pathway such as *Gja1* and *Acsf2* and the known corresponding intracellular signaling transductors *Lrp5* and *Lrp6* (Fig. 4e).

We further jointly analyzed CCC in mouse cortex datasets generated with three different technologies: Visium, seqFISH+ and STARmap. We found CCC patterns across the datasets that were consistent with existing knowledge, demonstrating the robustness of COMMOT (Extended Data Figs. 3–5). Details of the findings are given in the Supplementary Note. We also applied COMMOT to a large-scale spatial transcriptomics dataset, that is, Slide-seqV2 data of mouse hippocampus, containing expression of 23,264 genes in 53,173 beads (spatial spots), which are similar in size to individual cells[52] (Fig. 4f–h). Clustering based on CCC activities separated the spots into six clusters, of which clusters 1 and 2, consisting mostly of DentatePyramid, CA1_CA2_CA3_Subiculum, and interneuron cells, are generally active in CCC (Fig. 4f–h).

Signaling analysis in multi-cell resolution spatial transcriptomics data

Finally, we applied COMMOT to signaling analysis with Visium[16] spatial transcriptomics data, in which each spatial spot contains multiple cells. By analyzing the breast cancer data with 3,798 spots and 36,601 genes, we found clear spatial signaling directionality of midkine signaling, which was identified to be the most active (Fig. 5a), and the regions receiving such signals (Fig. 5b). To identify the genes that may be regulated by or regulate CCC, we used tradeSeq[53] to perform a differential expression test, in which the amount of received midkine signaling was used as the cofactor, analogous to a temporal differential expression test in which pseudotime is used as the cofactor (Fig. 5c,d). COL1A1 was identified as a significant positive differentially expressed gene with a distinct spatial pattern, whereas S100G was a significant negative differentially expressed gene with its own unique spatial pattern (Fig. 5c). Furthermore, as the received midkine signaling increases, the level of COL1A1 expression increases while the S100G expression level decreases (Fig. 5d). Adapting temporal differentially expressed gene analysis methods for scRNA-seq data to the signaling differentially expressed gene analysis of spatial transcriptomics data identifies relationships between gene expression and signaling activity, for example, between COL1A1 expression and midkine signaling. In general, good coverage of genes and a large number of cells or spots is preferred for CCC-associated differentially expressed gene analysis of spatial transcriptomics data.

Differential expression tests typically examine the pairwise correlation between a potential target gene and a cofactor. The higher-order interactions between multiple factors (multiple potential upstream genes and the cofactor) are often neglected. To prioritize the genes that are more likely to be regulated by CCC, we used a random forest model[54,55] in which the potential target gene is the output and the CCC cofactor and the top intracellular correlated genes are the input features. The feature importance of the cofactor in the trained model then served to quantify the unique information provided by the cofactor about the potential target gene, scoring the unique impact of individual ligand–receptor pairs on each of the identified signaling differentially expressed genes. This model showed that COL1A1 and S100G are distinctly impacted by various midkine ligand–receptor pairs (Fig. 5e). Such analysis may be carried out for any ligand–receptor pair expressed in the data, for example, the PD1 signaling pathway related to T-cell functions (Supplementary Fig. 20).

We also analyzed a Visium[16] dataset of mouse brain tissue with 3,355 spots and 32,285 genes (Fig. 5f,g). We found significant prosaposin signaling activity across the tissue (Fig. 5f), where broad protective roles of prosaposin in the nervous system were discovered[56], and fibroblast growth factor signaling was identified on the border of the cerebellar cortex (Fig. 5g), consistent with its known role in cerebellum patterning during development[57].

Robust identification of CCC direction and downstream target

To assess method robustness and efficiency we next studied the correlation between inferred CCC and the expression of known downstream genes, and compared COMMOT with three existing methods: CellChat[6], which was designed for scRNA-seq data, and Giotto[23] and CellPhoneDB v3[25], which were designed for spatial transcriptomics data.

To test robustness, we used the stage 6 *Drosophila* embryo, an extensively studied system[58,59]. An in situ spatial transcriptomic map was generated by integrating an scRNA-seq dataset with spatial single-cell resolution data[60] using SpaOTsc[31]. From subsampled data, COMMOT consistently identified CCC directions, cluster-level CCC and the signaling differentially expressed genes (Extended Data Fig. 6). See Methods for evaluation metrics and the Supplementary Note for more details.

Utilizing scSeqComm[61], a database of known target genes of ligand–receptor pairs combining major resources including Reactome, TTRUST and RegNetwork, we investigated the correlation between the inferred signaling activities and the expression of the corresponding target genes. We used three datasets analyzed in the previous sections with transcriptome or near-transcriptome gene coverage: Visium human breast cancer data, Visium mouse brain data and seqFISH+ mouse somatosensory cortex data. COMMOT was used to quantify all available ligand–receptor pairs in the CellChatDB. At the individual-spot scale, Spearman's correlation coefficient was computed for each ligand–receptor pair between the received signal and the average expression of the known downstream genes. The median correlations on the three datasets were 0.237, 0.180 and 0.230, respectively (Supplementary Fig. 21). At the cluster scale, we quantified the level of received signal using the average of the spots in the cluster.

We compared COMMOT with three methods that infer cluster-level CCC: CellChat[6], Giotto[23] and CellPhoneDB v3[25]. The activity of the downstream genes of a ligand–receptor pair was quantified as the percentage of significant positive differentially expressed genes of a cluster. By studying the correlation between the inferred CCC and the activity of known downstream genes, we found COMMOT to have a stronger correlation than the three methods for most datasets, and a comparable correlation to CellPhoneDB v3 in some cases (Supplementary Figs. 22–24). This evaluation can be further improved if more complete knowledge of gene regulation is available. With such a list, one may also formulate the evaluation as a classification problem. The differences between COMMOT and the three methods are illustrated in Supplementary Figs. 25–30 and discussed in the Supplementary Note. Furthermore, COMMOT can identify localized signaling hotspots compared with cluster-level approaches (Supplementary Figs. 31 and 32). For a specific ligand–receptor pair, COMMOT prioritizes regions containing its high signaling activity with low competition from other pairs (Supplementary Figs. 33 and 34), showing its unique strength.

To study algorithm efficiency, we found that COMMOT running time scales linearly with the number of non-zero elements in the CCC (Supplementary Fig. 35). The number of non-zero elements in the CCC

https://doi.org/10.1038/s41592-022-01728-4

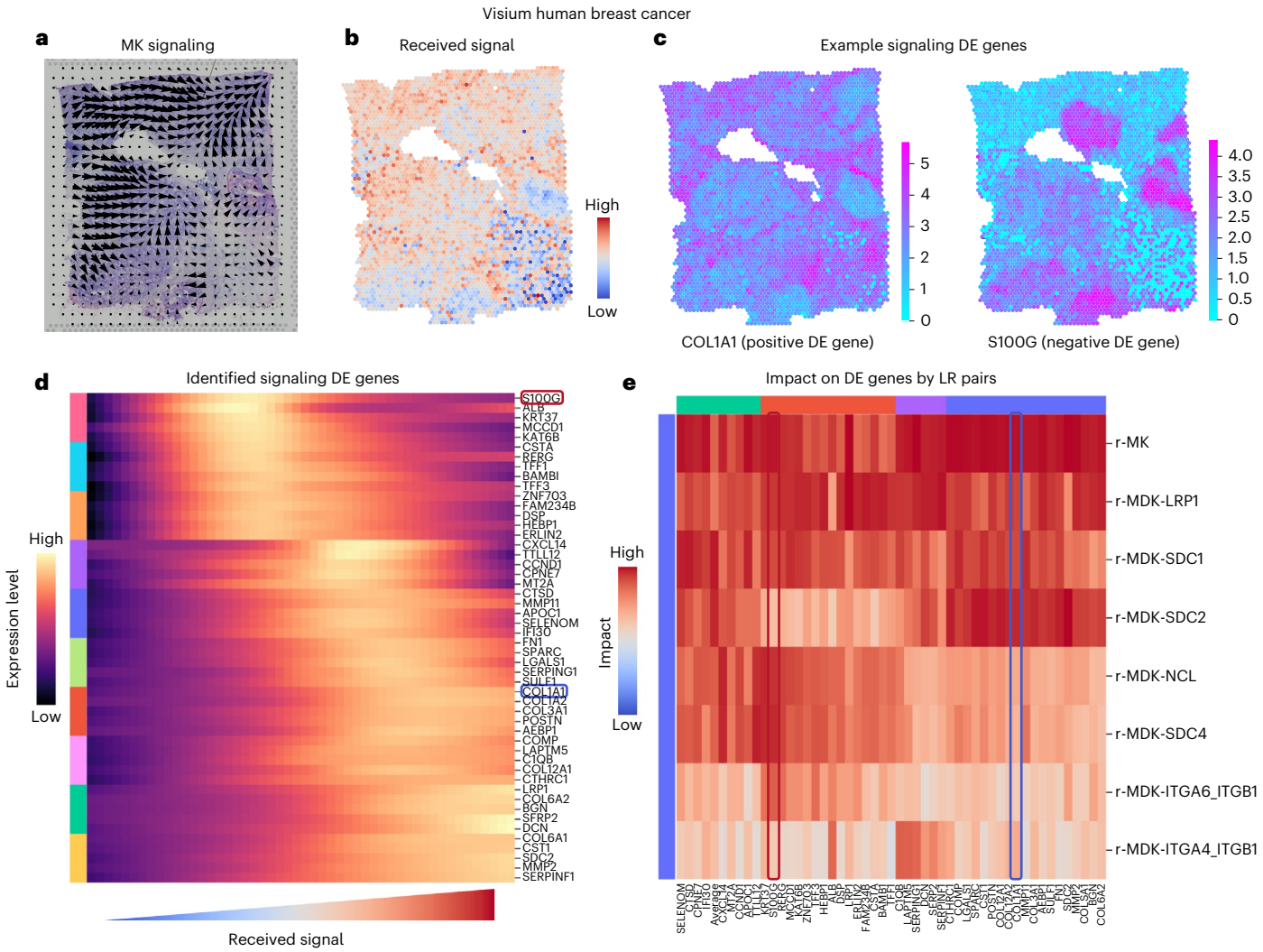


Fig. 5 | CCC inference using Visium spatial transcriptomics data. a–**e**, Midkine (MK) signaling in human breast cancer tissue. **a**, Spatial signaling direction. **b**, Amount of received signal by each spot. **c**, Two examples of differentially expressed (DE) genes due to signaling. **d**, Identification of the differentially expressed genes due to the total amount of received signal in the MK signaling pathway. **e**, Unique impact on the identified differentially expressed genes by the individual ligand–receptor pairs. **f**,**g**, Signaling in mouse brain tissue. The signaling direction (left) and the level of received signal (right) are shown for PSAP signaling (**f**) and FGF signaling (**g**).

matrices scales linearly with the number of locations in spatial transcriptomics data due to the spatial range constraint, and the memory usage also scales linearly with the number of locations given that only the finite values of the cost matrix and the non-zero values of the CCC matrix need to be stored. Thus, COMMOT can effectively handle the existing spatial transcriptomics datasets given that both computing time and memory usage both scale linearly with the number of spatial locations.

Discussion

To dissect CCC from the emerging spatial transcriptomics data we have developed COMMOT to infer CCC for all ligand and receptor species, simultaneously; visualize spatial CCC at various scales including a vector field visualization of spatial signaling directions; and analyze their downstream effects. This tool is based on collective optimal transport that incorporates both competing marginal distributions and

constrained transport plans, two important features that cannot be dealt with using current variants of optimal transport.

We have studied a wide range of data types with different spatial resolutions and gene coverage: in silico spatial transcriptomics data obtained by integrating scRNA-seq and spatial staining data, Visium, Slide-seq, STARmap, MERFISH and seqFISH+ spatial transcriptomics. COMMOT could consistently capture the CCC activities known from the literature. In human skin, COMMOT showed that higher WNT signaling increases the expression of several genes, a result confirmed by immunofluorescence staining. We acknowledge that false positives in our inferred CCC are inherently possible because spatial transcriptomics data do not directly represent protein abundancy and our method cannot capture protein-specific modifications such as protein phosphorylation, glycosylation, proteolytic cleavage into fragments, and dimerization, which certainly affect the signaling functions and, thus, the CCC mechanisms that COMMOT aims to infer. The reliability of CCC predictions is expected to significantly improve as emerging spatial proteomics approaches mature.

The spatial distance constraint used to capture the effect of ligand diffusivity is usually determined by several factors, including protein weight and tortuosity of extracellular space[62]. It is difficult to accurately estimate this parameter for every pair in the database. In our model the local short-range interactions are emphasized even when the spatial distance range is increased (Supplementary Fig. 36). Thus, when screening many ligand–receptor pairs a uniform and relatively large spatial distance limit may be used to avoid missing important interactions. Once the important interactions are identified, an accurate estimation of this parameter would further refine the prediction to remove false-positive CCC links.

Most recently, several methods and packages have been introduced to study CCC with spatial transcriptomics data. SpatialDM[63] evaluates the co-expression of ligand and receptor genes; SpaTalk[64] and stMLnet[65] are focused on signaling target genes; HoloNet[66] studies the joint impact from different combinations of CCC events; and DeepLinc[67] constructs de novo cell–cell interaction landscapes without the need for annotated ligand and receptor genes. Although COMMOT has a different focus, these methods arguably complement each other when studying different aspects of CCC.

With the foreseeable availability of temporal sequences of spatial transcriptomics data[68], CCC dynamics may be elucidated, for example by extending collective optimal transport into a dynamic optimal transport formulation. The PDE model of CCC can be generalized to further incorporate the intracellular gene regulatory network. While traditional optimal transport is powerful at integrating a pair of datasets and multimarginal optimal transport[69] integrates multiple datasets, the collective optimal transport is able to effectively control the coupling and deal with competing species, which is useful for a broad range of problems beyond CCC inference.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01728-4.

References

1. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
2. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88 (2021).
3. Almet, A. A., Cang, Z., Jin, S. & Nie, Q. The landscape of cell–cell communication through single-cell transcriptomics. *Curr. Opin. Syst. Biol.* **26**, 12–23 (2021).
4. Türei, D. et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.* **17**, e9923 (2021).
5. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
6. Jin, S. et al. Inference and analysis of cell–cell communication using CellChat. *Nat. Commun.* **12**, 1088 (2021).
7. Noël, F. et al. Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nat. Commun.* **12**, 1089 (2021).
8. Wang, S., Karikomi, M., Maclean, A. L. & Nie, Q. Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Res.* **47**, e66 (2019).
9. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2020).
10. Hu, Y., Peng, T., Gao, L. & Tan, K. CytoTalk: de novo construction of signal transduction networks using single-cell transcriptomic data. *Sci. Adv.* **7**, eabf1356 (2021).
11. Tsuyuzaki, K., Ishii, M. & Nikaido, I. Uncovering hypergraphs of cell–cell interaction from single cell RNA-sequencing data. Preprint at https://doi.org/10.1101/566182 (2019).
12. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563**, 347–353 (2018).
13. Abbasi, S. et al. Distinct regulatory programs control the latent regenerative potential of dermal fibroblasts during wound healing. *Cell Stem Cell* **27**, 396–412 (2020).
14. Armingol, E. et al. Inferring a spatial code of cell–cell interactions across a whole animal body. *PLoS Comput. Biol.* **18**, e1010715 (2022).
15. Dries, R. et al. Advances in spatial transcriptomic data analysis. *Genome Res.* **31**, 1706–1718 (2021).
16. Ståhl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
17. Rodriques, S. G. et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
18. Eng, C.-H. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239 (2019).
19. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
20. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
21. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
22. Palla, G. et al. Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).
23. Dries, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 78 (2021).
24. Pham, D. T. et al. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell–cell interactions and spatial trajectories within undissociated tissues. Preprint at https://doi.org/10.1101/2020.05.31.125658 (2020).
25. Garcia-Alonso, L. et al. Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro. *Nat. Genet.* **53**, 1698–1711 (2021).

26. Arnol, D., Schapiro, D., Bodenmiller, B., Saez-Rodriguez, J. & Stegle, O. Modeling cell–cell interactions from spatial molecular data with spatial variance component analysis. *Cell Rep.* **29**, 202–211 (2019).
27. Tanevski, J., Flores, R. O. R., Gabor, A., Schapiro, D. & Saez-Rodriguez, J. Explainable multiview framework for dissecting spatial relationships from highly multiplexed data. *Genome Biol.* **23**, 97 (2022).
28. Fischer, D. S., Schaar, A. C. & Theis, F. J. Modeling intercellular communication in tissues using spatial graphs of cells. *Nat. Biotechnol.* (2022).
29. Forrow, A. et al. Statistical optimal transport via factored couplings. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (eds. Chaudhuri, K. & Sugiyama, M.) 89 2454–2465 (PMLR, 2019).
30. Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943 (2019).
31. Cang, Z. & Nie, Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat. Commun.* **11**, 2084 (2020).
32. Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. Gene expression cartography. *Nature* **576**, 132–137 (2019).
33. Peyré, G. & Cuturi, M. Computational optimal transport: with applications to data science. *Foundations and Trends in Machine Learning* **11**, 355–607 (2019).
34. Villani, C. *Optimal Transport: Old and New* (Springer Science & Business Media, 2008).
35. Ramilowski, J. A. et al. A draft network of ligand–receptor-mediated multicellular signalling in human. *Nat. Commun.* **6**, 7866 (2015).
36. Figalli, A. The optimal partial transport problem. *Arch. Rational Mech. Anal.* **195**, 533–560 (2010).
37. Bonneel, N. & Coeurjolly, D. SPOT: sliced partial optimal transport. *ACM Transactions on Graphics* **38**, 89 (2019).
38. Chizat, L., Peyré, G., Schmitzer, B. & Vialard, F.-X. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation* **87**, 2563–2609 (2018).
39. Wang, S. et al. Single cell transcriptomics of human epidermis identifies basal stem cell transition states. *Nat. Commun.* **11**, 4239 (2020).
40. Choi, Y. S. et al. Distinct functions for Wnt/β-catenin in hair follicle stem cell proliferation and survival and interfollicular epidermal homeostasis. *Cell Stem Cell* **13**, 720–733 (2013).
41. Bamberger, C. et al. Activin controls skin morphogenesis and wound repair predominantly via stromal cells and in a concentration-dependent manner via keratinocytes. *Am. J. Pathol.* **167**, 733–747 (2005).
42. Mou, H. et al. Dual SMAD signaling inhibition enables long-term expansion of diverse epithelial basal cells. *Cell Stem Cell* **19**, 217–231 (2016).
43. Moffitt, J. R. et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
44. Froemke, R. C. & Young, L. J. Oxytocin, neural plasticity, and social behavior. *Annu. Rev. Neurosci.* **44**, 359–381 (2021).
45. Warfvinge, K., Krause, D. & Edvinsson, L. The distribution of oxytocin and the oxytocin receptor in rat brain: relation to regions active in migraine. *J. Headache Pain* **21**, 10 (2020).
46. He, Y. et al. ClusterMap for multi-scale clustering analysis of spatial gene expression. *Nat. Commun.* **12**, 5909 (2021).
47. Bie, C. et al. Insulin-like growth factor 1 receptor drives hepatocellular carcinoma growth and invasion by activating Stat3-Midkine-Stat3 loop. *Dig. Dis. Sci.* **67**, 569–584 (2022).
48. Sandovici, I. et al. The imprinted Igf2–Igf2r axis is critical for matching placental microvasculature expansion to fetal growth. *Dev. Cell* **57**, 63–79 (2022).
49. Marchese, M. J., Li, S., Liu, B., Zhang, J. J. & Feng, L. Perfluoroalkyl substance exposure and the BDNF pathway in the placental trophoblast. *Front. Endocrinol. (Lausanne)* **12**, 694885 (2021).
50. Jeyarajah, M. J., Jaju Bhattad, G., Kops, B. F. & Renaud, S. J. Syndecan-4 regulates extravillous trophoblast migration by coordinating protein kinase C activation. *Sci. Rep.* **9**, 10175 (2019).
51. Bocchi, R. et al. Perturbed Wnt signaling leads to neuronal migration delay, altered interhemispheric connections and impaired social behavior. *Nat. Commun.* **8**, 1158 (2017).
52. Stickels, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).
53. Van den Berge, K. et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201 (2020).
54. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
55. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
56. Meyer, R. C., Giddens, M. M., Coleman, B. M. & Hall, R. A. The protective role of prosaposin and its receptors in the nervous system. *Brain Res.* **1585**, 1–12 (2014).
57. Yaguchi, Y. et al. Fibroblast growth factor (FGF) gene expression in the developing cerebellum suggests multiple roles for FGF signaling during cerebellar morphogenesis and development. *Dev. Dyn.* **238**, 2058–2072 (2009).
58. Lécuyer, E. et al. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* **131**, 174–187 (2007).
59. Tomancak, P. et al. Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **8**, R145 (2007).
60. Karaiskos, N. et al. The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
61. Baruzzo, G., Cesaro, G. & Di Camillo, B. Identify, quantify and characterize cellular communication from single-cell RNA sequencing data with scSeqComm. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btac036 (2022).
62. Lander, A. D., Nie, Q. & Wan, F. Y. M. Do morphogen gradients arise by diffusion? *Dev. Cell* **2**, 785–796 (2002).
63. Li, Z., Wang, T., Liu, P. & Huang, Y. SpatialDM: Rapid identification of spatially co-expressed ligand-receptor reveals cell–cell communication patterns. Preprint at https://doi.org/10.1101/2022.08.19.504616 (2022).
64. Shao, X. et al. Knowledge-graph-based cell–cell communication inference for spatially resolved transcriptomic data with SpaTalk. *Nat. Commun.* **13**, 4429 (2022).
65. Cheng, J., Yan, L., Nie, Q. & Sun, X. Modeling spatial intercellular communication and multilayer signaling regulations using stMLnet. Preprint at https://doi.org/10.1101/2022.06.27.497696 (2022).
66. Li, H., Ma, T., Hao, M., Wei, L. & Zhang, X. Decoding functional cell–cell communication events by multi-view graph learning on spatial transcriptomics. Preprint at https://doi.org/10.1101/2022.06.22.496105 (2022).
67. Li, R. & Yang, X. De novo reconstruction of cell interaction landscapes from single-cell spatial transcriptome data with DeepLinc. *Genome Biol.* **23**, 124 (2022).
68. Longo, S. K., Guo, M. G., Ji, A. L. & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat. Rev. Genet.* **22**, 627–644 (2021).
69. Pass, B. Multi-marginal optimal transport: theory and applications. *ESAIM Math. Model. Numer. Anal.* **49**, 1771–1790 (2015).

Methods

Full details of the theoretical background and implementation of COMMOT can be found in the Supplementary Information.

COMMOT model

COMMOT constructs a collection of CCC networks through various predefined ligand–receptor pairs (user-defined or from aggregated ligand–receptor interaction databases) by solving a global optimization problem that accounts for potential higher-order interactions between the multiple ligand and receptor species. To this end, we introduce collective optimal transport that determines a collection of optimal transport plans for all pairs of species that can be coupled simultaneously. As a result, the coupling between a species pair will affect other couplings and vice versa, which cannot be realized in traditional optimal transport[34]. The collective optimal transport results in a large-scale optimization problem for which new algorithms are needed, and thus we present one based on the efficient Sinkhorn iteration[70].

For a spatial transcriptomics dataset of n_s spatial locations and a set of n_l ligand species and n_r receptor species, a collective optimal transport problem is formulated as follows:

$$\min_{\boldsymbol{P}\in\Gamma} \sum_{(i,j)\in I} \langle \boldsymbol{P}_{i,j,\cdot,\cdot}, \boldsymbol{C}_{(i,j)} \rangle_F + \sum_i F(\mu_i) + \sum_j F(\nu_j),$$

$$\Gamma = \left\{ \boldsymbol{P} \in \mathbb{R}_+^{n_l \times n_r \times n_s \times n_s} : \boldsymbol{P}_{i,j,\cdot,\cdot} = \boldsymbol{0} \text{ for } (i,j) \notin I, \sum_{j,l} \boldsymbol{P}_{i,j,k,l} \le \boldsymbol{X}^L_{i,k}, \sum_{i,k} \boldsymbol{P}_{i,j,k,l} \le \boldsymbol{X}^R_{j,l} \right\},$$

$$\mu_i(k) = \boldsymbol{X}^L_{i,k} - \sum_{j,l} \boldsymbol{P}_{i,j,k,l},\ \nu_j(l) = \boldsymbol{X}^R_{j,l} - \sum_{i,k} \boldsymbol{P}_{i,j,k,l} \tag{1}$$

where $\boldsymbol{X}^L_{i,k}$ is the expression level of ligand i on spot k, $\boldsymbol{X}^R_{j,l}$ is the expression level of receptor j on spot l and F penalizes the untransported mass $\boldsymbol{\mu}_i$ and $\boldsymbol{\nu}_j$. The coupling matrix $\boldsymbol{P}_{i,j,k,l}$ scores the signaling strength from spot k to spot l through the pair consisting of the ligand i and receptor j for $(i,j) \in I$ where I is the index set of ligand and receptor species that can bind. The cost matrix $\mathbf{C}_{(i,j)}$ is based on the thresholded distance matrix such that its kl-th entry equals $\varphi(\mathbf{D}_{k,l})$ if $\mathbf{D}_{k,l} \le T_{(i,j)}$ and infinity otherwise, where $\mathbf{D}$ is the Euclidean distance matrix for the distances between the spots, $T_{(i,j)}$ is the spatial limit of signaling through the pair of ligand i and receptor j, and φ is a scaling function, such as square or exponential. When the ligands or receptors contain heteromeric units, the minimum of units is used by default in the package to represent the amount of ligand or receptor. For example, if receptor species j is composed of two subunits, the minimum of them in spot l is used to represent the level of this receptor species $\mathbf{X}^R_{j,l}$.

Collective optimal transport algorithm

To solve the collective optimal transport problem described above, we rewrite the original problem as:

$$\min_{\hat{\boldsymbol{P}},\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}}\ge 0} \langle \hat{\boldsymbol{P}}, \hat{\boldsymbol{C}} \rangle_F + \epsilon_p H(\hat{\boldsymbol{P}}) + \epsilon_\mu H(\hat{\boldsymbol{\mu}}) + \epsilon_\nu H(\hat{\boldsymbol{\nu}}) + \rho\left(\|\hat{\boldsymbol{\mu}}\|_1 + \|\hat{\boldsymbol{\nu}}\|_1\right),$$
$$\text{s.t.}\, \hat{\boldsymbol{P}}\boldsymbol{1}^n = \boldsymbol{a} - \hat{\boldsymbol{\mu}},\ \hat{\boldsymbol{P}}^T\boldsymbol{1}^m = \boldsymbol{b} - \hat{\boldsymbol{\nu}} \tag{2}$$

where $\hat{\boldsymbol{P}}$ is obtained by reshaping $\boldsymbol{P}$ such that $\hat{\boldsymbol{P}}_{(i-1)\times n_s+k,(j-1)\times n_s+l} = \boldsymbol{P}_{i,j,k,l}$. The cost matrix $\hat{\boldsymbol{C}}$ is obtained similarly and we set $\hat{\boldsymbol{C}}_{(i-1)\times n_s+k,(j-1)\times n_s+l} = \infty$ for ligand i and receptor j that cannot bind. The marginal distributions are constructed such that $\boldsymbol{a}_{(i-1)\times n_s+k} = \boldsymbol{X}^L_{i,k}$ and $\boldsymbol{b}_{(j-1)\times n_s+l} = \boldsymbol{X}^R_{j,l}$. Entropy regularization is added to speed up computation and smooth the result with $H(\boldsymbol{x}) = \sum_i x_i(\ln(x_i) - 1)$.

When the entropy regularization terms have the same coefficient values, $\epsilon = \epsilon_p = \epsilon_\mu = \epsilon_\nu$, the problem can be efficiently solved with a stabilized Sinkhorn iteraction[70]

$$\boldsymbol{f}^{(l+1)} \leftarrow \epsilon \log \boldsymbol{a} + \boldsymbol{f}^{(l)} - \epsilon \log\left(e^{\frac{\boldsymbol{f}^{(l)}}{\epsilon}} \odot e^{-\frac{\boldsymbol{C}}{\epsilon}} e^{\frac{\boldsymbol{g}^{(l)}}{\epsilon}} + e^{\frac{\boldsymbol{f}^{(l)}-\rho}{\epsilon}}\right),$$
$$\boldsymbol{g}^{(l+1)} \leftarrow \epsilon \log \boldsymbol{b} + \boldsymbol{g}^{(l)} - \epsilon \log\left(e^{\frac{\boldsymbol{g}^{(l)}}{\epsilon}} \odot e^{-\frac{\boldsymbol{C}^T}{\epsilon}} e^{\frac{\boldsymbol{f}^{(l+1)}}{\epsilon}} + e^{\frac{\boldsymbol{g}^{(l)}-\rho}{\epsilon}}\right), \tag{3}$$

for $l \ge 0$ with arbitrary initial $\boldsymbol{f}^{(0)}$ and $\boldsymbol{g}^{(0)}$. The resulting numerical solution to the optimization problem can be constructed by $\hat{\boldsymbol{P}}^* = e^{(\boldsymbol{f}\oplus\boldsymbol{g}-\boldsymbol{C})/\epsilon}$. The formulation in Eq. (2) solved by the algorithm in Eq. (3) was used to generate the results in this study. The derivation of the algorithm, and that of algorithms for the general case in which the regularization terms have different coefficients, is described in the Supplementary Information.

Spatial signaling direction

To visualize the spatial signaling directions, we estimate a spatial vector field $\boldsymbol{V} \in \mathbb{R}^{n_s \times d}$ of signaling directions given a CCC matrix $\boldsymbol{S} \in \mathbb{R}_+^{n_s \times n_s}$ obtained from collective optimal transport algorithm where $\boldsymbol{S}_{i,j}$ is the strength of the signal sent by spot i to spot j. The ith row of $\boldsymbol{V}$ represents the spatial signaling direction. We construct two vector fields, $\boldsymbol{V}^s$ and $\boldsymbol{V}^r$ describing the direction to/from which the spots are sending/receiving signals, respectively. Specifically, $\boldsymbol{V}^s_i = (\sum_j \boldsymbol{S}_{i,j}) \times \mathcal{N}\left(\sum_{j\in N_i^s} \boldsymbol{S}_{i,j}\mathcal{N}(\boldsymbol{x}_j - \boldsymbol{x}_i)\right)$, where $\mathcal{N}(\mathbf{x}) = \mathbf{x}\|\mathbf{x}$ and N_i^s is the index set of top k signal-sending spots with the largest value on the ith row of $\mathbf{S}$. Similarly, $\boldsymbol{V}^r_i = (\sum_j \boldsymbol{S}_{j,i}) \times \mathcal{N}\left(\sum_{j\in N_i^r} \boldsymbol{S}_{j,i}\mathcal{N}(\boldsymbol{x}_i - \boldsymbol{x}_j)\right)$, where N_i^r is the index set of top k signal-receiving spots with the largest value on the ith column of $\mathbf{S}$.

Cluster-level CCC

To elucidate CCC among cell states or local groups of spots, we aggregate the spot-by-spot CCC matrix $\boldsymbol{S}$ to a cluster-by-cluster matrix $\boldsymbol{S}^{cl}$. The signaling strength from cluster i to cluster j is quantified as $\boldsymbol{S}^{cl}_{i,j} = \sum_{(k,l)\in I^{cl}_{i,j}} \boldsymbol{S}_{k,l}/|I^{cl}_{i,j}|$, where $I^{cl}_{i,j} = \{(k,l) : L_k = i, L_l = j\}$ and L_k is the cluster label of spot k. The significance (P value) of the cluster-level CCC is determined by performing n independent permutations of the cluster labels and computing the percentile of the original signaling strength in the signaling strengths resulting from these label permutations. Permuting cluster labels after computing the spot-level CCC matrices may neglect communications between different clusters. To address this limitation, we provide an option that randomly permutes the locations of all spots or the spots within each cluster and then computes the spot-level CCC matrices.

Evaluation metrics

The spatial signaling direction is described by a vector field defined on a discretized tissue space consisting of n grid points and is represented by an array $\boldsymbol{V} \in \mathbb{R}^{n\times d}$. The cosine distance is used to compare the vector field $\boldsymbol{V}_{\text{sub}}$ from subsampled data with the one from the full data $\boldsymbol{V}_{\text{full}}$ and is defined as

$$d_{\cos}(\boldsymbol{V}_{\text{full}}, \boldsymbol{V}_{\text{sub}}) = \sum_i \|\boldsymbol{V}_{\text{full}}(i)\| \left[1 - \boldsymbol{V}_{\text{full}}(i) \cdot \boldsymbol{V}_{\text{sub}}(i) / (\|\boldsymbol{V}_{\text{full}}(i)\| \|\boldsymbol{V}_{\text{sub}}(i)\|)\right] / \sum_i \|\boldsymbol{V}_{\text{full}}(i)\|.$$

To compare two cluster-level CCC networks $\boldsymbol{S}^{cl}_1$ and $\boldsymbol{S}^{cl}_2$, we first binarize them such that the edges with $P < 0.05$ are kept in the edge sets $\bar{\boldsymbol{S}}^{cl}_1$ and $\bar{\boldsymbol{S}}^{cl}_2$. Then, the Jaccard distance is used for quantitative comparison, $d_{\text{Jaccard}}\left(\bar{\boldsymbol{S}}^{cl}_1, \bar{\boldsymbol{S}}^{cl}_2\right) = 1 - \left|\bar{\boldsymbol{S}}^{cl}_1 \cap \bar{\boldsymbol{S}}^{cl}_2\right| / \left|\bar{\boldsymbol{S}}^{cl}_1 \cup \bar{\boldsymbol{S}}^{cl}_2\right|$.

The Spearman's correlation coefficient is used to quantify the correlation between the inferred signaling activity and the activity of the known target genes across the cell clusters, defined as $\text{cov}\left(\text{R}\left(\boldsymbol{X}^{LR}\right), \text{R}\left(\boldsymbol{X}^{tgt}\right)\right) / \left(\sigma_{\text{R}(\boldsymbol{X}^{LR})}\sigma_{\text{R}(\boldsymbol{X}^{tgt})}\right)$, where $\boldsymbol{X}^{LR}_i$ is the average received signal through a ligand–receptor pair in cell cluster i, and $\boldsymbol{X}^{tgt}_i$ is the activity of the known target genes of this ligand–receptor pair in cell

cluster i quantified as the percentage of differentially expressed genes. The function R converts the vectors into ranks and σ is the standard deviation of the rank variables.

Downstream gene analysis

After computing the CCC matrix $\boldsymbol{S}$ of a ligand–receptor pair or a signaling pathway, genes that are potential downstream targets of the corresponding CCC can be identified. The amount of signal received by each spot is quantified by $\boldsymbol{r} \in \mathbb{R}^{n_s}$ where $\boldsymbol{r}_i = \sum_j \boldsymbol{S}_{j,i}$. Then the tradeSeq package[53] is used to identify the genes that are differentially expressed with respect to $\boldsymbol{r}$, which we call differentially expressed CCC genes.

The identified differentially expressed CCC genes may be regulated by other genes in cells through gene regulation. To further prioritize the downstream genes, the expressions of which are affected by CCC, we train a random forest regression model[54,55] that takes a potential downstream gene as the output, and $\boldsymbol{r}$ and a collection of highly correlated genes as input features. The unique impact of CCC on this potential downstream gene is quantified by the feature importance (Gini importance computed as the mean of total impurity decrease in each tree) of $\boldsymbol{r}$ in the trained random forest model. The inclusion of highly correlated genes in a cell as input features emphasizes the amount of information of potential target genes explained by inferred CCC, which is unlikely to be explained only by intracellular interactions. If such a dilution of importance is not preferred, the users may choose a smaller number of highly correlated genes as input features. The implementation in the scikit-learn package[55] is used.

CellChat, Giotto and CellPhoneDB analysis

For the CellChat analysis the spatial data were treated as non-spatial scRNA-seq data, and the count matrix was first normalized using the normalizeData function. The data were then filtered using the functions identifyOverExpressedGenes and identifyOverExpressedInteractions with the default parameters. The cluster-level communication scores in CellChat were computed using the computeCommunProb function with default parameters, and the results were further filtered using the filterCommunication function with min.cells set to 10. The ligand–receptor pairs categorized under 'Secreted Signaling' in the CellChatDB were examined. For Giotto analysis, the count data were first normalized using the normalizeGiotto function with default parameters. A spatial network was then created using the createSpatialNetwork function with the k-nearest neighbors method and k set to 100 and the maximum distance threshold of 1000 µm for Visium data and 500 µm for seqFISH+ data. The heteromeric ligand–receptor pairs in CellChatDB were converted to pairs of individual subunits. The spatCellCellcom function was then used to generate the cluster-level communication scores with the adjust_method set to fdr. For CellPhoneDB v3 analysis, the distance between clusters was quantified as the average distance between cells from the pair of clusters. The command 'cellphonedb method statistical_analysis' was used to generate CellPhoneDB results with the threshold parameter set to 0.1.

Immunostaining and fluorescence in situ hybridization

Frozen tissue sections (10 µm) were fixed with 4% paraformaldehyde in PBS for 15 min. Ten percent BSA in PBS was used for blocking. Following blocking, 5% BSA and 0.1% Triton X-100 in PBS was used for permeabilization. The following antibodies were used: mouse anti-KRT5 (1:100; Santa Cruz Biotechnology, sc-32721), mouse anti-KRT15 (1:100; Santa Cruz Biotechnology, sc-47697), mouse anti-BCAM (1:100; Santa Cruz Biotechnology, sc-365191), mouse anti-FGF7 (1:100; Santa Cruz Biotechnology, sc-365440), mouse anti-STMN1 (1:100; Santa Cruz Biotechnology, sc-48362); mouse anti-IGFBP6 (1:500; Abgent, AP6764b); mouse anti-PMAIP1 (1:100; Santa Cruz Biotechnology, sc-56169), mouse anti-POSTN (1:100; Santa Cruz Biotechnology, sc-398631); mouse anti-FLG (1:100; Santa Cruz Biotechnology, sc-66192); rabbit anti-LOR (1:1000; abcam, ab85679); mouse anti-TYRO3 (1:100; LSBio, LS-C114523-100); rabbit anti-GAS6 (1:100; abcam, ab227174); and rabbit anti-PROS1 (1:100; Proteintech, 16910-1-AP). Secondary antibodies include Alexa Fluor 488 (1:500; Jackson ImmunoResearch, 715-545-150, 711-545-152) and Cy3 AffiniPure (1:500; Jackson ImmunoResearch, 711-165-152, 111-165-003). Slides were mounted with Prolong Diamond Antifade Mountant containing DAPI (Molecular Probes). Confocal images were acquired at room temperature (22.2 °C) on a Zeiss LSM700 laser scanning microscope with a Plan-Apochromat ×20 objective or ×40 and ×63 oil immersion objectives.

Frozen neonatal human foreskin tissue sections were used for RNA in situ hybridization using RNAscope kit v2 (323100, Advanced Cell Diagnostics) as per the manufacturer's instructions. The following *Homo sapiens* probes from Advanced Cell Diagnostics were used: Tyro3 probe (429611), Gas6 (427811-C2) and Pros1 (506991-C2). Confocal images were acquired at room temperature on an Olympus FV3000 confocal microscope with a Plan-Apochromat ×20 objective or ×40 and ×60 oil immersion objectives.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The original public data used in this work can be accessed through the following links: *Drosophila* embryo spatial and scRNA-seq data: Dream Single cell Transcriptomics Challenge through Synapse ID (syn15665609)[60]; human epidermal scRNA-seq data[39]: GEO accession code GSE147482 (protocols involving human skin data were approved by the Institutional Review Board of the University of California, Irvine); mouse hypothalamic preoptic region MERFISH data[43]: original data available at Dryad[71] at the link https://doi.org/10.5061/dryad.8t8s248 (this work used the preprocessed data through the Squidpy package[22] with the utility squidpy.datasets.merfish); mouse placenta STARmap data[46]: downloaded from Code Ocean (https://codeocean.com/capsule/9820099/tree/v1) with the https://doi.org/10.24433/CO.6072400.v1; mouse brain STARmap data[20]: processed data were downloaded from the same repository as the mouse placenta STARmap data; mouse somatosensory cortex seqFISH+ data[18]: downloaded through the Giotto package[23]; mouse hippocampus Slide-seqV2 data[52]: downloaded from the Broad Institute Single Cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP815/sensitive-spatial-genome-wide-expression-profiling-at-cellular-resolution#study-summary); breast cancer Visium data: downloaded from the 10X Genomics website (https://www.10xgenomics.com/resources/datasets/human-breast-cancer-block-a-section-1-1-standard-1-1-0); mouse brain (sagittal posterior) Visium data: downloaded from the 10X Genomics website (https://www.10xgenomics.com/resources/datasets/mouse-brain-serial-section-1-sagittal-anterior-1-standard-1-1-0). The ligand–receptor pairs with secreted ligands, as categorized in the CellChatDB[6], were used and can be accessed at http://www.cellchat.org/cellchatdb/. The downstream target genes were taken from scSeqComm[61] and the target gene libraries TF_TG_TRRUSTv2 and TF_TG_TRRUSTv2_RegNetwork_High_mouse were used for human and mouse, respectively.

Code availability

The open-source software is available at https://github.com/zcang/COMMOT. The code for reproducing the presented analysis results is available at https://doi.org/10.5281/zenodo.7272562 (ref. [72]).

References

70. Cuturi, M. Sinkhorn distances: lightspeed computation of optimal transportation distances. *Adv. Neural Inf. Processing Syst.* **26**, 2292–2300 (2013).

71. Moffitt, J. R. et al. Data from: Molecular, spatial and functional single-cell profiling of the hypothalamic preoptic region. *Dryad, Dataset*, https://doi.org/10.5061/dryad.8t8s248 (2018).
72. Cang, Z. et al. COMMOT: Screening cell–cell communication in spatial transcriptomics via collective optimal transport (0.0.2). *Zenodo* https://doi.org/10.5281/zenodo.7272562 (2022).

Acknowledgements

This work was supported by two National Science Foundation (NSF) grants (DMS1763272 and CBET2134916), a grant from the Simons Foundation (594598 to Q.N.), a Chan Zuckerberg Initiative grant (AN-0000000062) and three National Institutes of Health grants (U01AR073159, R01DE030565 and R01AR079150). Z.C.'s work was partially supported by a startup grant from North Carolina State University and an NSF grant (DMS2151934). Y.Z.'s work was supported by a grant from the Simons Foundation through Grant No. 357963 and NSF grant DMS2142500. Z.C. thanks W. Zhao at University of California, Irvine for helpful discussions.

Author contributions

Z.C., Y.Z., and Q.N. conceived the method. Z.C. implemented the method. Z.C. and A.A.A. generated the numerical results. R.R., A.S. and S.X.A. generated the experimental results. Z.C., R.R., A.S., M.V.P., S.X.A. and Q.N. interpreted the results, generated the diagrams and wrote the paper. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data are available for this paper at https://doi.org/10.1038/s41592-022-01728-4.

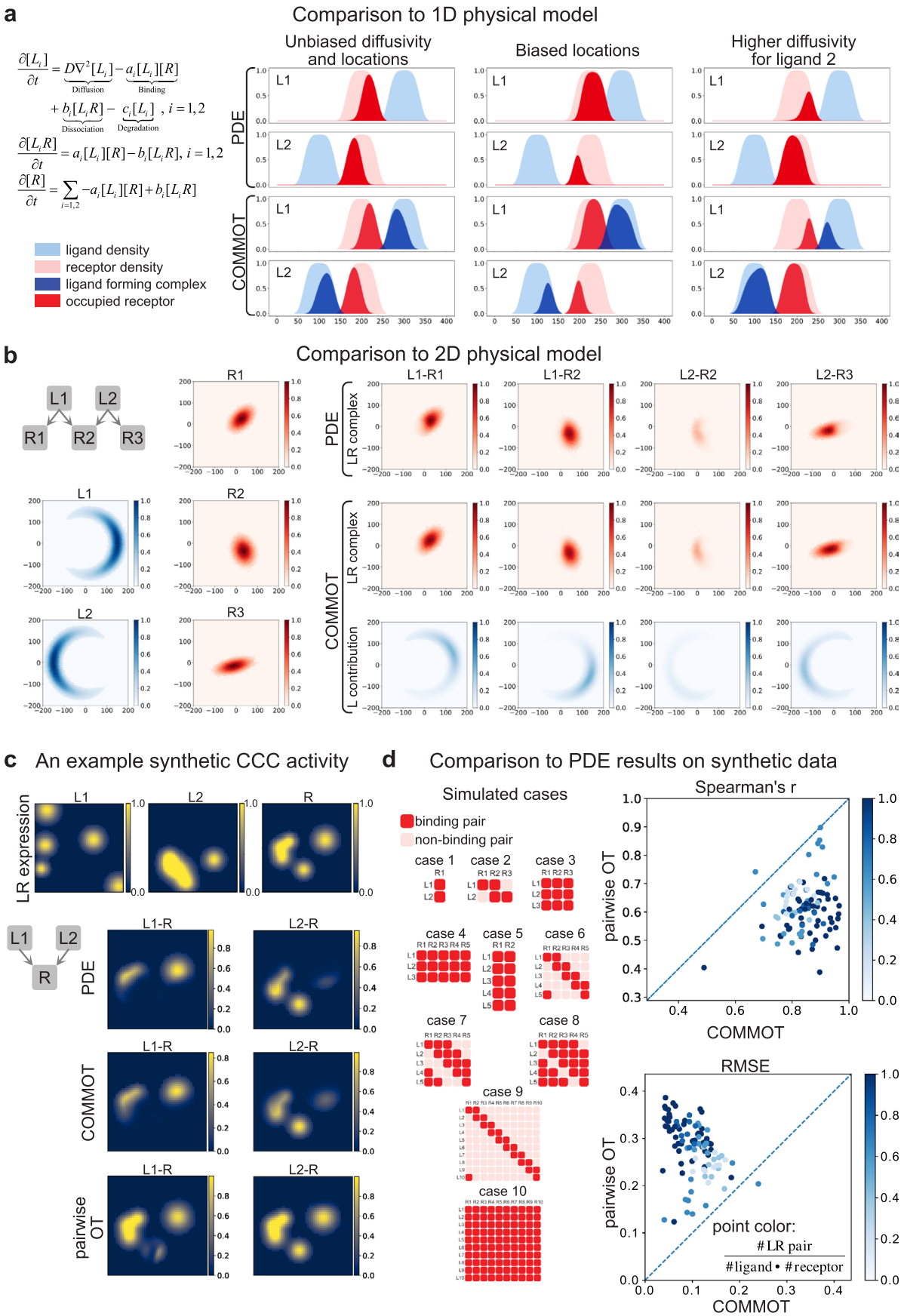
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01728-4.

Correspondence and requests for materials should be addressed to Qing Nie.

Peer review information *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Lei Tang, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41592-022-01728-4

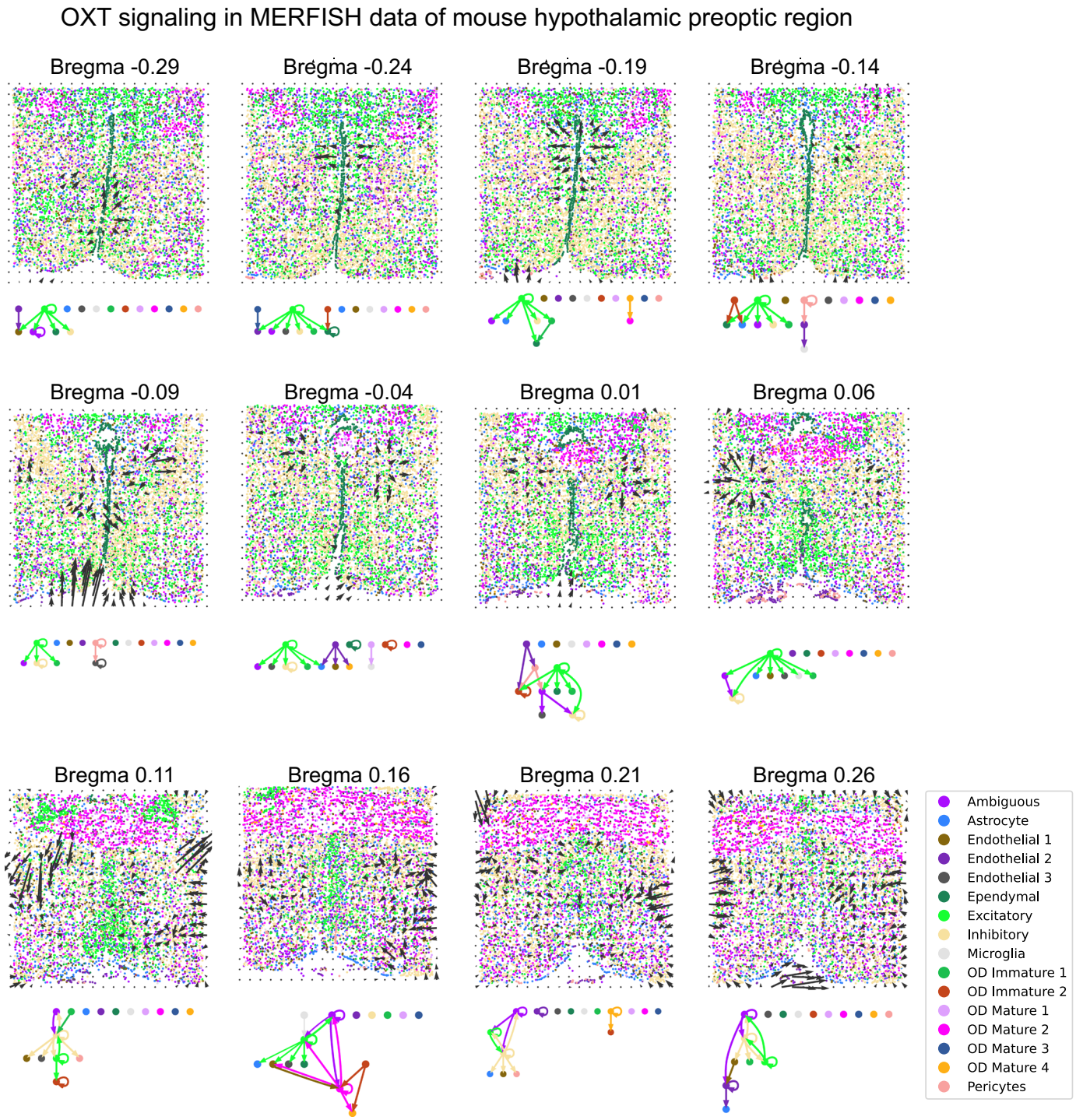


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41592-022-01728-4

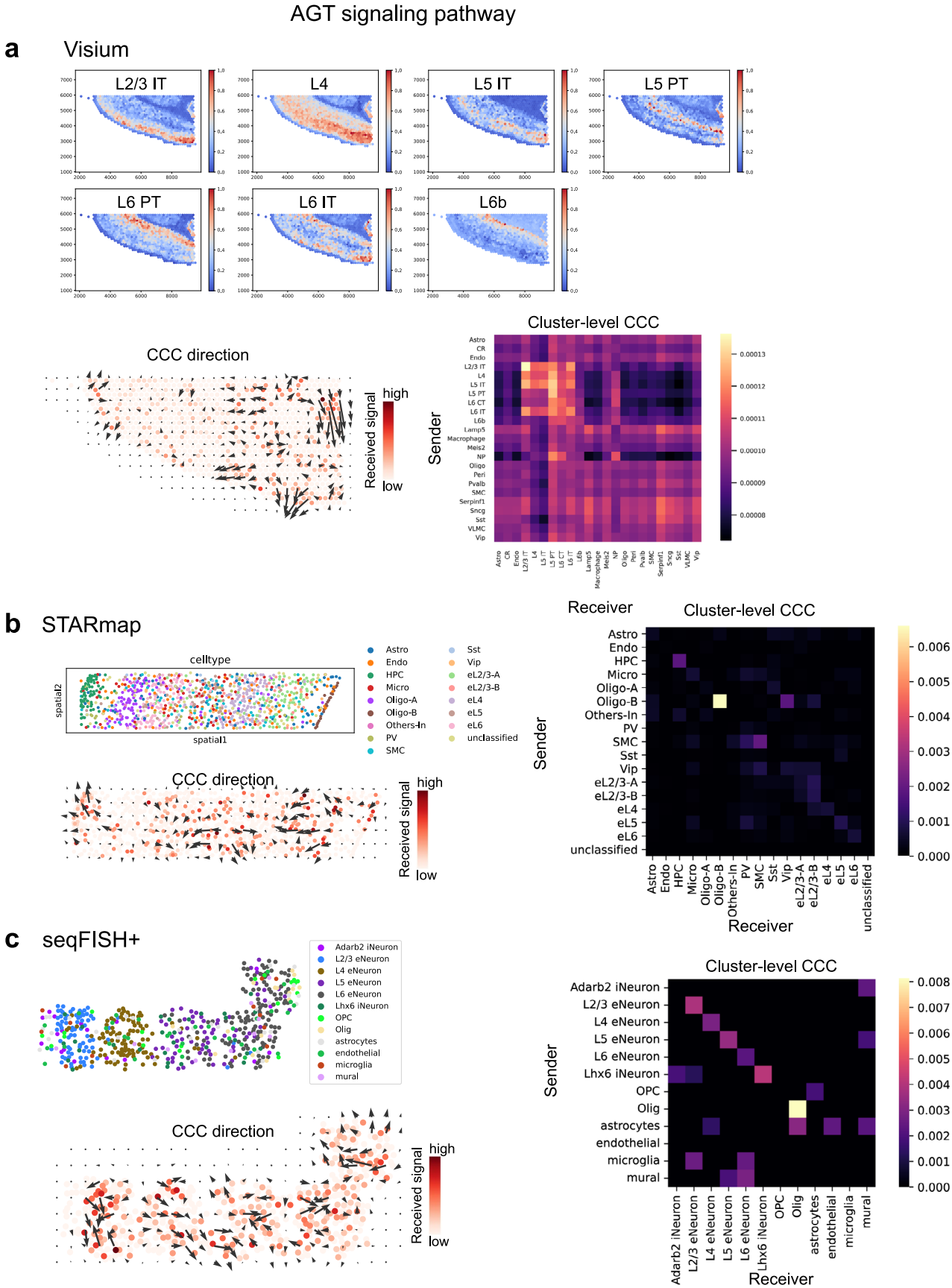
Extended Data Fig. 1 | Validation using simulated data by partial differential equations (PDE) model. The example PDE model where two ligand species can bind to the same receptor. The inference by COMMOT is compared to the simulation results in several 1-dimensional cases. **b** Comparison to simulated results in a 2-dimensional case with three ligand species and two receptor species. **c** An example of randomly generated 2-dimensional benchmark with two ligand species that binds to the same receptor. The simulated result, inference by COMMOT, and inference by pairwise method are shown. **d** Ten different cases of ligand–receptor binding and the performance of COMMOT and pairwise OT (with the same spatial limit as COMMOT but each LR pair examined separately) obtained by comparing to simulated results.

https://doi.org/10.1038/s41592-022-01728-4



Extended Data Fig. 2 | OXT CCC in MERFISH mouse hypothalamic preoptic region. The inferred signaling directions and cluster-level CCC of OXT signaling in each of the slice of the MERFISH data.

https://doi.org/10.1038/s41592-022-01728-4

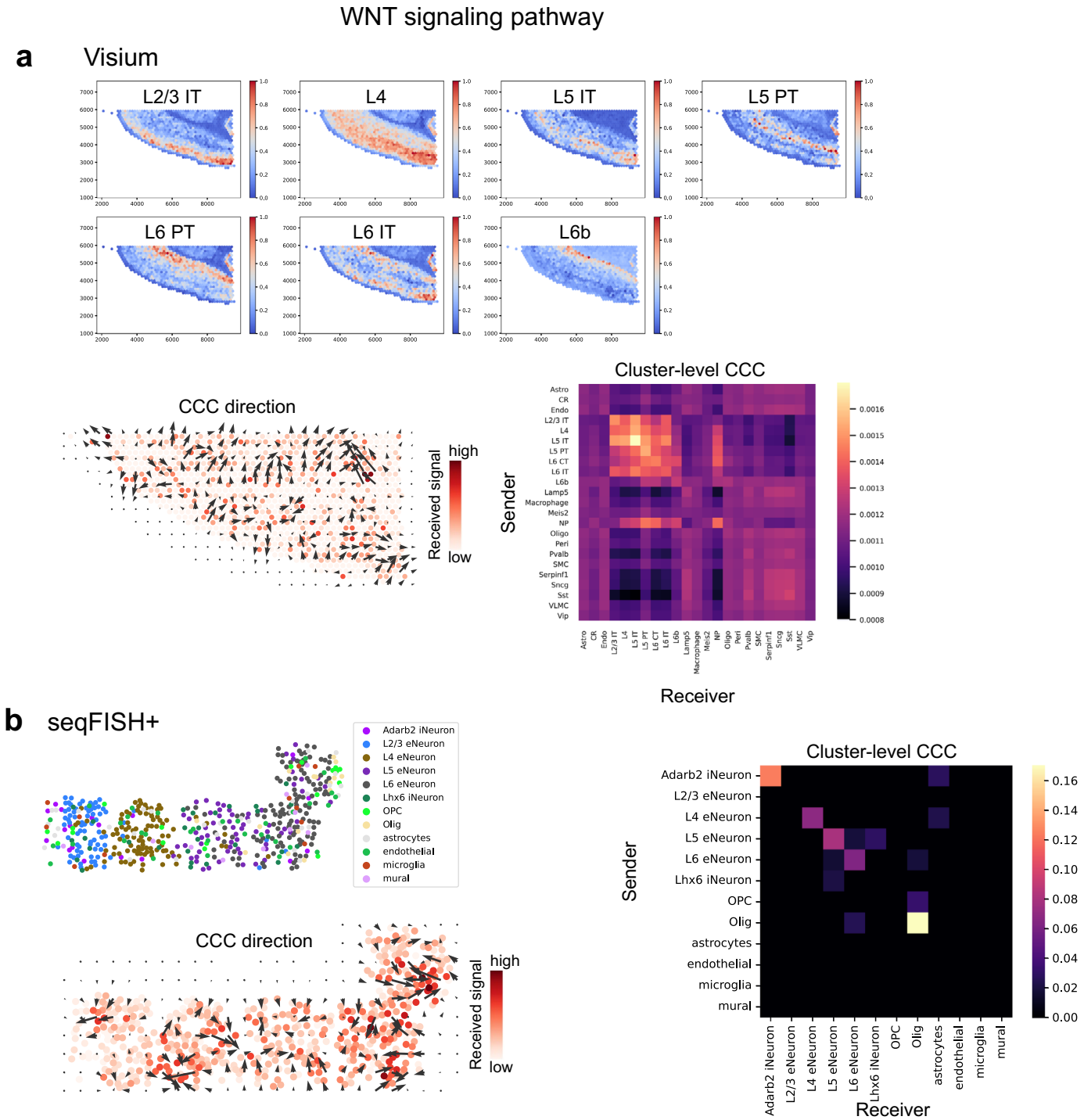


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41592-022-01728-4

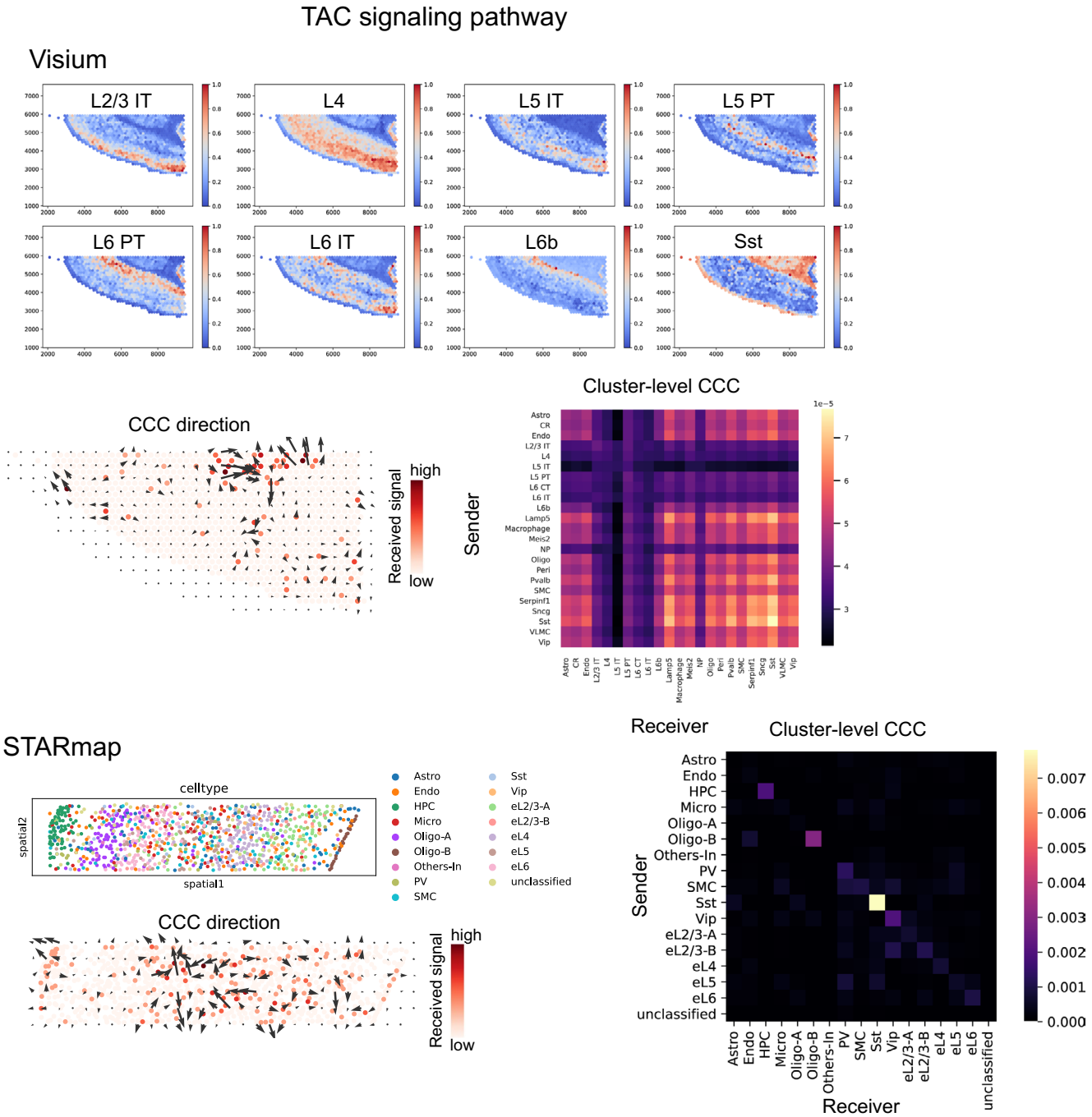
Extended Data Fig. 3 | AGT signaling pathway in mouse cortex. 1) Cell type plots, 2) spatial directions of CCC, and 3) heatmaps of cluster-level CCC of the AGT signaling pathway in **a** Visium, **b** STARmap, and **c** seqFISH+ mouse cortex data. Across these three datasets, AGT signaling was identified in neurons. Spatially, neurons in the L2-3 region were identified as strong receivers of AGT ligands across the three datasets. Interestingly, a striped signaling pattern was observed, wherein strong signals within individual layers form stripes, while weak signals form inter-stripe regions. Strong AGT signaling activity among oligodendrocytes was also identified in both STARmap and seqFISH+ datasets.

https://doi.org/10.1038/s41592-022-01728-4



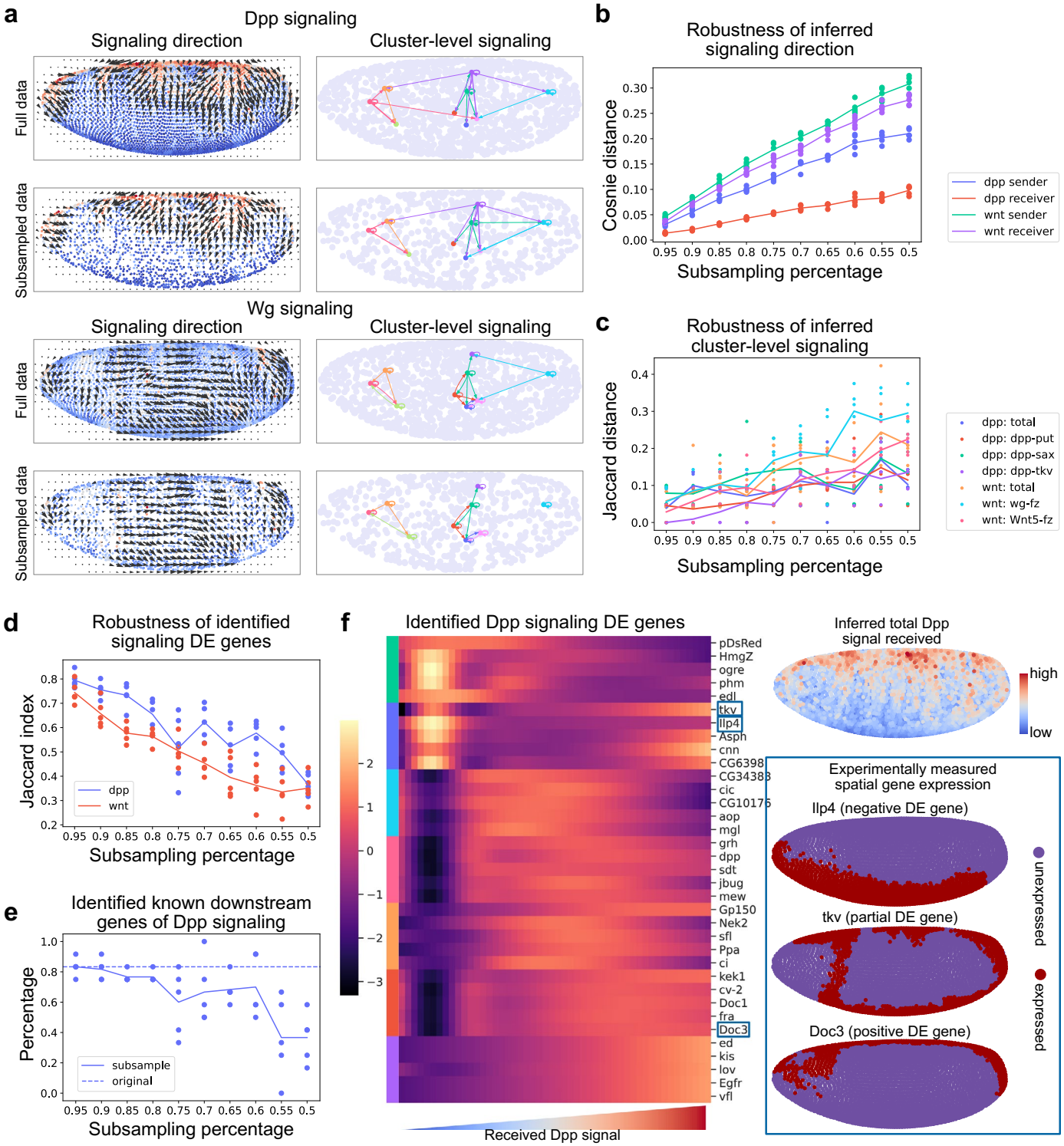
Extended Data Fig. 4 | WNT signaling pathway in mouse cortex. 1) Cell type plots, 2) spatial directions of CCC, and 3) heatmaps of cluster-level CCC of the WNT signaling pathway in **a** Visium and **b** seqFISH+ mouse cortex data. In both Visium and seqFISH+ cortex datasets, we inferred WNT signaling to be active across different cortical layers. In both datasets, we identified WNT signaling to be relatively low in layer 5, compared to other layers.

https://doi.org/10.1038/s41592-022-01728-4



Extended Data Fig. 5 | TAC signaling pathway in mouse cortex. 1) Cell type plots, 2) spatial directions of CCC, and 3) heatmaps of cluster-level CCC of the TAC signaling pathway in **a** Visium and **b** STARmap mouse cortex data. TAC (tachykinin neuropeptide family) signaling activity was consistently found in both Visium and STARmap cortex datasets to be active in non-neuronal cells and in inhibitory neurons, especially in somatostatin-expressing neurons (Sst).

https://doi.org/10.1038/s41592-022-01728-4



Extended Data Fig. 6 | Robustness of CCC analysis on a well-studied drosophila embryo dataset. a Spatial signaling direction and signaling among cell clusters for Dpp and Wg signaling pathways. **b** Robustness of inferred signaling direction evaluated by comparing the direction obtained from subsampled dataset to the one from the full dataset using cosine distance. Each point is an independent test and the line shows the average of the tests. **c** Robustness of inferred cluster-level communication evaluated by comparing random subsamples to the full dataset using the Jaccard distance. **d** Robustness of downstream gene identification. **e** Percentage of known downstream genes that are identified as differentially expressed gene due to signaling activity. **f** Examples of the identified positively, negatively, and partially differentially expressed genes associated to Dpp signaling. For panels b–e, the averages of 5 independent random subsampling are plotted.

nature portfolio

Corresponding author(s): Qing Nie

Last updated by author(s): Oct 31, 2022

Reporting Summary

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

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	The open-source software has been uploaded to Github (https://github.com/zcang/COMMOT) and analysis code and results will be uploaded to Zenodo upon publication. The following versions were used for the software mentioned in the manuscript: CellChat version 1.1.3, Giotto version 1.0.3, CellPhoneDB version 3.1.0, tradeSeq version 1.0.1, scikit-learn version 1.0.2.

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The original public data used in this work can be accessed through the following links: (1) Drosophila embryo spatial and scRNA-seq data: Dream Single cell Transcriptomics Challenge through Synapse ID (syn15665609) [Karaiskos, N. et al., Science, 2017]. (2) Human epidermal scRNA-seq data [Wang, S. et al., Nat.

Commun., 2020]: GEO assession codes, GSE147482. (3) Mouse hypothalamic preoptic region MERFISH data [Moffitt, J. R. et al., Science, 2018]: original data available at Dryad[Moffitt, J. R. et al., Dryad, Dataset, 2018] through the link https://doi.org/10.5061/dryad.8t8s248. This work used the preprocessed data through the Squidpy package [Palla, G. et al., Nat. Methods, 2022] with the utility squidpy.datasets.merfish. (4) Mouse placenta STARmap data [He, Y. et al., Nat. Commun, 2021]: downloaded from Code Ocean (https://codeocean.com/capsule/9820099/tree/v1) with DOI: 10.24433/CO.6072400.v1. (5) Mouse brain STARmap data [Wang, X. et al., Science, 2018]: the processed data was downloaded from the same repository as the mouse placenta STARmap data. (6) Mouse somatosensory cortex seqFISH+ data [Eng, C.-H. L. et al., Nature, 2019]: downloaded through Giotto package [Dries, R. et al., Genome Biol., 2021]. (7) Mouse hippocampus Slide-seqV2 data [Stickels, R. R. et al., Nat. Biotechnol., 2020]: downloaded from Broad Institute Single Cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP815/sensitive-spatial-genome-wide-expression-profiling-at-cellular-resolution#study-summary). (8) Breast cancer Visium data: downloaded from 10X Genomics website (https://www.10xgenomics.com/resources/datasets/human-breast-cancer-block-a-section-1-1-standard-1-1-0). (9) Mouse brain (sagittal posterior) Visium data: downloaded from 10X Genomics website (https://www.10xgenomics.com/resources/datasets/mouse-brain-serial-section-1-sagittal-anterior-1-standard-1-1-0).

The ligand-receptor pairs with secreted ligand based on CellChatDB database [Jin, S. et al., Nat. Commun., 2021] were used and can be accessed at http://www.cellchat.org/cellchatdb/. The downstream target genes were taken from scSeqComm [Baruzzo, G. et a., Bioinformatics, 2022] and the target gene libraries named TF_TG_TRRUSTv2 and TF_TG_TRRUSTv2_RegNetwork_High_mouse were used for human and mouse respectively.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	All tissue samples were archived discarded and de-identified neonatal foreskins not collected for the purpose of this study. All tissues are from males due to the nature of the tissue. No other covariants were used in the collection of the tissues.
Population characteristics	All human samples were from archived tissue not collected for the purpose of this study and our group was blinded to all characteristics of human subjects.
Recruitment	Not applicable. Tissue was collected as discarded and de-identified samples from available newborns.
Ethics oversight	Institutional Review Board of the University of California, Irvine.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	No sample size calculations were performed for experiments. Discarded and de-identified human foreskin samples were used for immunohistochemical analysis with at least 3 biological replicates. Biological replicate sample size and size sufficiency was chosen due to the similarity of immunohistochemical staining.
Data exclusions	No data was excluded from the analysis.
Replication	All experiments were reproduced a minimum of three times.
Randomization	No randomization was necessary for immunohistochemical analysis because all samples were used to describe pathway status.
Blinding	Single cell RNA-seq and spatial transcriptomic analyses were unbiased. All cells were analyzed using computational algorithms that were not biased to recognize any particular cell types. All available tissue were used for immunohistochemical analysis. As only wild-type tissue was used with no manipulations, no blinding was necessary.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Antibodies

Antibodies used

The following antibodies were used: mouse anti-KRT5 (1:100; Santa Cruz Biotechnilogy; sc-32721), mouse anti-KRT15 (1:100; Santa Cruz Biotechnology; sc-47697), mouse anti-BCAM (1:100; Santa Cruz Biotechnology; sc-365191), mouse anti-FGF7 (1:100; Santa Cruz Biotechnology; sc-365440), mouse anti-STMN1 (1:100; Santa Cruz Biotechnology; sc-48362); mouse anti-IGFBP6 (1:500; Abgent; AP6764b); mouse anti-PMAIP1 (1:100; Santa Cruz Biotechnology; sc-56169), mouse anti-POSTN (1:100; Santa Cruz Biotechnology; sc-398631); mouse anti-FLG (1:100; Santa Cruz Biotechnology; sc-66192); rabbit anti-LOR (1:1000; abcam; ab85679); mouse anti-TYRO3 (1:100; LSBio; LS-C114523-100); rabbit anti-GAS6 (1:100; abcam; ab227174); and rabbit anti-PROS1 (1:100; Proteintech; 16910-1-AP). Secondary antibodies include Cy3 AffiniPure (1:500; Jackson ImmunoResearch; 711-165-152, 111-165-003).

Validation

mouse anti-KRT5 (1:100; Santa Cruz Biotechnilogy; sc-32721): Sung JS et al. 2020. Oncogene. 39(3):664-676.
mouse anti-KRT15 (1:100; Santa Cruz Biotechnology; sc-47697): Busslinger GA et al. 2021. Cell Rep. 34(10):108819.
mouse anti-BCAM (1:100; Santa Cruz Biotechnology; sc-365191): Zhao J et al. 2022. Clin Epigenetics. 14(1):99.
mouse anti-FGF7 (1:100; Santa Cruz Biotechnology; sc-365440): Chen X et al. 2022. Br J Pharmacol. 179(5):1102-1121.
mouse anti-STMN1 (1:100; Santa Cruz Biotechnology; sc-48362): Hu Z et al. 2020. Cancer Cell. 37(2):226-242.e7.
mouse anti-IGFBP6 (1:500; Abgent; AP6764b): manufacturer validated with human samples via western blot and IHC. Synthetic peptide of human IFGBP6 used as an antigen.
mouse anti-PMAIP1 (1:100; Santa Cruz Biotechnology; sc-56169): Palanikumar L et al. 2021. Nat Commun. 12(1):3962.
mouse anti-POSTN (1:100; Santa Cruz Biotechnology; sc-398631): Mircea M et al. 2021. Genome Biol. 23(1):18.
mouse anti-FLG (1:100; Santa Cruz Biotechnology; sc-66192): Dai X et al. 2022. J Invest Dermatol. 142:136-144.e3.
rabbit anti-LOR (1:1000; abcam; ab85679): Zhou Q et al. 2021. J Invest Dermatol. 141:152-163.
mouse anti-TYRO3 (1:100; LSBio; LS-C114523-100): manufacturer validated with human samples via western blot and IHC. Full length recombinant protein of human TYRO3 used as an antigen.
rabbit anti-GAS6 (1:100; abcam; ab227174): manufacturer validated with human samples via western blot and IHC. Recombinant fragment protein of human GAS6 used as an antigen.
rabbit anti-PROS1 (1:100; Proteintech; 16910-1-AP): Wang ZH et al. 2015. Mol Med Rep. 12(3):3279-3284.